The vocabulary growth of EFL learners in Saudi Arabia: The role of individual differences, digital flashcard learning and quiz frequency

by

Abdullah Albalawi

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of English Language and Linguistics

College of Arts and Law

University of Birmingham

July 2024

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT

Despite the substantial expansion in vocabulary research since the 1980s (Laufer, 2009; Meara, 2002), we still know very little about how vocabulary develops over time and what factors influence this development (Pellicer-Sánchez, 2019; Webb & Nation, 2017). The first study of the thesis aimed to address this by examining the vocabulary breadth growth of EFL learners over a school semester (12 weeks). It measured the vocabulary growth (meaning recognition and meaning recall) of 141 Saudi intermediate school (aged 15) and secondary school (aged 16) students using the Updated Vocabulary Levels Test (Webb et al., 2017). To explain the expected variation in vocabulary growth, the study examined the role of individual differences focusing on three key factors: out-of-class exposure (e.g., watching TV and playing video games), self-regulation and motivation. The main finding from this study is that vocabulary growth in an EFL context can be low and slow (Nurweni & Read, 1999; Siyanova-Chanturia & Webb, 2016; Webb & Chang, 2012), and after many years of school instruction, students might still not develop a good knowledge of even the highest frequency vocabulary (i.e., the most frequent 1000 word-families). Additionally, out-of-class exposure and motivation were significant predictors of vocabulary learning.

The second study aimed to address the low knowledge of high frequency vocabulary found in the first study. Given the limited time of many EFL classes, it employed digital flashcard learning in out-of-class settings and included in-class quizzes to make sure that students genuinely engage with vocabulary learning and potentially benefit from the testing effect (Karpicke & Roediger, 2007). However, it was unclear based on the previous research how frequently quizzes should occur for optimal vocabulary learning. The second study aimed

to address this gap by first examining the effect of quizzing (quiz vs. no-quiz) followed by an examination of the effect of quiz frequency (weekly, biweekly and monthly) on vocabulary learning over a school semester (eight weeks). Secondary school students (n = 76, age = 16-17) learned 120 target words using digital flashcards in naturalistic out-of-class settings using their personal devices. The second study had two main findings. First, the groups who received quizzes showed significant vocabulary improvement on the posttest while the group who did not receive quizzes did not make any significant vocabulary gains. This finding suggests that supplementing out-of-class vocabulary learning with in-class quizzes can be an effective vocabulary learning approach. It also suggests that students' willingness to engage in out-of-class language learning (i.e., extra-curricular learning) should not be taken for granted when there is no source of external motivation (Seibert Hanson & Brown, 2019). Second, there were no significant differences in the learning gains between the three quiz frequency groups (weekly, biweekly and monthly), suggesting that more frequent quizzes do not necessarily lead to more vocabulary learning.

The thesis overall makes valuable contributions to both vocabulary theory and practice. The first study enhances our understanding of the nature of vocabulary knowledge by examining vocabulary growth longitudinally while taking into account the role of individual differences. The second study offers practical recommendations to help language learners learn vocabulary more effectively. The two studies combined make important strides in advancing L2 vocabulary learning, instruction and research.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to Allah, the Almighty, for providing me with the strength, knowledge, and perseverance to complete this thesis. Without His blessings, none of this would have been possible.

I am deeply grateful to my supervisors, Dr. Gareth Carrol and Dr. Petra Schoofs, for their invaluable guidance and support throughout my doctoral research. Their insightful feedback and encouragement have been instrumental in shaping this thesis.

I would like to thank the examiners, Dr. Beatriz Gonzalez-Fernandez and Dr. Jing Huang, for their time and valuable feedback during the viva. Your insights have been greatly appreciated.

I would like to acknowledge the financial support provided by the Saudi Electronic University, which enabled me to conduct my research effectively.

To my parents and family, I owe an immense debt of gratitude. Their belief in my abilities, love and encouragement sustained me through the challenges of this PhD journey. Their sacrifices and understanding are deeply appreciated.

I extend my heartfelt thanks to my beloved wife for her unwavering support, patience and encouragement throughout this journey. Her love and understanding have been a constant source of strength and motivation for me.

Lastly, I am grateful to my dear friends Mohammed Albalawi and Nife Alanazi for their invaluable assistance with data collection. Your help has been crucial to the success of this research, and I am thankful for your support and dedication.

Parts of this thesis have been accepted for publication or under review in peer reviewed journals:

The role of individual differences in vocabulary learning (Chapter 2, section 2.3)

Albalawi, A. (2024). The role of individual differences in L2 vocabulary learning: A review of out-of-class exposure, strategic learning and motivation. Australian Journal of Applied Linguistics. 7(2), 1-20. https://doi.org/10.29140/ajal.v7n2.1641

Measuring vocabulary growth (Chapter 2, section 2.2)

Albalawi, A. (2024). Key issues and considerations in measuring vocabulary growth: A methodological review. Vocabulary Learning and Instruction, 13(2), 1-13. https://doi.org/10.29140/vli.v13n2.1604

Vocabulary learning in Saudi Arabia (Chapter 3)

Albalawi, A. (under review). A review of English vocabulary learning, instruction and research in Saudi Arabia.

Study 1 (Chapter 4)

Albalawi, A., Carrol, G. & Schoofs, P. (under review). The vocabulary growth of EFL learners in Saudi Arabia and the role of out-of-class exposure, self-regulation and motivation.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

LIST OF ABBREVIATIONS

EFL	English as a Foreign Language
L1	First Language
L2	Second Language
LLSs	Language Learning Strategies
SLA	Second Language Acquisition
UVLT	Updated Vocabulary Levels Test
VLT	Vocabulary Levels Test
VLSs	Vocabulary Learning Strategies
VST	Vocabulary Size Test

1. Introduction

Vocabulary is a key component of language. Its significance is captured by Wilkins's often cited sentence "without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (Wilkins, 1972, p. 110). Learners themselves are aware of the importance of vocabulary and regard lexical errors to be more serious than all other types of errors (Politzer, 1978). While vocabulary received relatively little attention in the past, as noted by Meara (1980), the research landscape has undergone significant transformation in the years that followed his article, giving rise to a substantial body of work (Laufer, 2009; Meara, 2002; Milton & Hopwood, 2022). This is evident in the multitude of research topics investigated within vocabulary and the publication of several comprehensive books summarizing research in this area (Durrant et al., 2022; Nation, 2022; Schmitt & Schmitt, 2020; Webb, 2019).

Research on vocabulary can be broadly categorized into four key research areas: vocabulary development, vocabulary testing, vocabulary and psycholinguistics and vocabulary and corpus linguistics (Durrant et al., 2022). Vocabulary development research focuses on the teaching and learning of vocabulary. It investigates topics such as how vocabulary knowledge is conceptualized (Anderson & Freebody, 1981; Daller et al., 2007; Henriksen, 1999; Melka, 1997; Nation, 1990; Richards, 1976), how vocabulary develops over time (Schmitt, 1998; Webb & Chang, 2012) and the role of incidental and intentional learning in vocabulary learning and instruction (Huckin & Coady, 1999; Hulstijn, 2001,

2003; Hulstijn et al., 1996; Laufer & Hulstijn, 2001; Thomas, 2020). Vocabulary testing focuses on the development and evaluation of vocabulary knowledge measures. Topics investigated in this area include vocabulary test development (Laufer & Nation, 1999; Nation, 2012; Webb et al., 2017), validation (Aviad-Levitzky et al., 2019; Beglar, 2010) and evaluation (Milton, 2009; Read, 2019; Schmitt et al., 2019). Psycholinguistic research on vocabulary focuses on the mental processes involved in vocabulary comprehension and production. It employs on-line methods such as eye-tracking and lexical decision tasks to uncover how vocabulary is learned and processed (Ellis & Beaton, 1993; Nakata & Elgort, 2021; Pellicer-Sánchez, 2016; A. Wang & Pellicer-Sánchez, 2022). Corpus linguistics research within vocabulary examines how large bodies of text can inform vocabulary learning and assessment. It focuses on topics such as the development of wordlists which can be used to guide the selection of the most useful words to learn (Coxhead, 2000; Nation, 2016; West, 1953) and the analysis of vocabulary in textbooks (Alsaif & Milton, 2012; Sun & Dang, 2020).

Despite this expansion, studies examining vocabulary growth using longitudinal data are scarce (Pellicer-Sánchez, 2019; Webb & Nation, 2017). Webb and Nation (2017, p. 68) state that "Surprisingly, there are relatively few studies of L2 vocabulary growth; and questions such as ‘How many words should be learned per week/per year/during a course?’ remain unanswered". Schmitt (2010, p. 156) emphasizes that "vocabulary learning is longitudinal and incremental in nature, and only research designs with a longitudinal element can truly describe it”. Schmitt (2019) also surveyed what vocabulary areas need more research and received 36 suggestions from 23 vocabulary scholars. The most frequently mentioned area in need of more research was longitudinal studies. The same

point has been reiterated by Pellicer-Sanchez (2019) in a paper calling for more longitudinal research on vocabulary growth. The shortage of longitudinal studies is not limited to vocabulary acquisition research, but also extends to the field of second language acquisition as a whole (Dörnyei, 2007).

The present thesis therefore aimed to investigate vocabulary development over time, and in particular the factors that can make this more or less successful. Two longitudinal studies were designed to assess this. The first study aimed to examine the vocabulary growth of Saudi intermediate and secondary EFL students over a school semester. It took into account key individual differences that might affect vocabulary growth. The three factors examined were motivation, self-regulation and out-of-class exposure, each of which has been suggested to be key in the literature (Sundqvist, 2009, 2024; Tseng et al., 2006; Tseng & Schmitt, 2008). The second study built on this by investigating how we can help low level EFL learners with limited vocabulary knowledge expand their vocabulary. Given the limited time available for foreign language instruction in most contexts, it examined the effectiveness of out-of-class vocabulary learning from digital flashcards in naturalistic settings. The contribution of the second study lies in investigating the effect of quizzing (quizzes vs. no quizzes) and quiz frequency (weekly, biweekly, monthly) on vocabulary learning over a school semester while taking into account key individual difference factors (motivation and self-regulation). The combination of these two studies goes some way to filling the gap in the literature regarding longitudinal research on vocabulary learning, and contributes to our understanding of the complex and dynamic set of factors that affect vocabulary development in this context. The organization of the thesis is discussed in the following section, which also provides a brief description of the six thesis chapters.

1.1 Organization of the thesis

The thesis begins by providing a general literature review of the key and relevant concepts in vocabulary research (Chapter 2) and reviews research on vocabulary in Saudi Arabia (Chapter 3). These two background chapters are followed by two empirical studies (Chapters 4 and 5). Finally, the thesis concludes with a general discussion and conclusion chapter that synthesizes the findings from the two studies and relates them to previous research (Chapter 6).

The first section in the general literature review chapter (section 2.1) provides an introduction to the key ideas and concepts in second language vocabulary research. The section begins by defining the concept of a word and how it has been operationalized in vocabulary research. It also discusses the recently debated issue of unit of counting (i.e., lemma, flemma and word-families) and how it affects vocabulary instruction, assessment and research. Additionally, the section introduces the important role of frequency in vocabulary learning and highlights the fact that some words are more useful than others because they are more frequent. The third section discusses how the construct of vocabulary knowledge has been defined and conceptualized in the literature. Section four highlights the importance of vocabulary in the four language skills (listening, reading, speaking and writing) and in overall language proficiency. Finally, it describes the two complementary methods of vocabulary learning (incidental and intentional) and explains the importance of both for well-developed vocabulary knowledge.

Section 2.2 starts by discussing fundamental concepts in vocabulary assessment such as validity and reliability and discusses key issues and considerations in vocabulary testing.

It also critically examines common vocabulary breadth tests and highlights their strengths and weaknesses. The section also considers the limitations of cross-sectional data when measuring vocabulary growth and highlights the importance of using longitudinal data. Finally, it concludes by reviewing studies that have examined vocabulary growth from a longitudinal perspective.

The role of individual differences in language and vocabulary learning is discussed in section 2.3. It starts by briefly discussing the role of individual differences in second language and vocabulary learning. It then moves on to discuss in separate subsections the role of out-of-class exposure, strategic learning and motivation in vocabulary learning.

Chapter 3 provides a review of vocabulary research in Saudi Arabia. It focuses on vocabulary learning, instruction and research. Vocabulary learning addresses the vocabulary knowledge of Saudi EFL students and their mastery of high frequency words. Vocabulary instruction focuses on how teachers in Saudi Arabia approach vocabulary learning and teaching. Finally, vocabulary research critically examines the research conducted to improve Saudi EFL students' vocabulary learning.

Chapter 4 represents the first study in the thesis, which examines the vocabulary growth of Saudi EFL students over a school semester. Additionally, it investigates the role of individual differences, in particular the role of out-of-class exposure, strategic vocabulary learning and motivation in vocabulary learning. The study aimed to improve our understanding of how vocabulary knowledge (meaning recognition and meaning recall) develops over time and how growth is influenced by individual differences. The

pedagogical implications of the study for vocabulary learning in an EFL context are discussed, as well as their implications for English instruction in Saudi Arabia.

Chapter 5 presents the second study in the thesis, which suggests one way of helping low level language learners improve their vocabulary knowledge. The study focuses on the use of digital flashcards in out-of-class settings to efficiently enhance and boost learners' vocabulary learning. It first investigates whether supplementing out-of-class digital flashcard learning with in-class quizzes can lead to more effective vocabulary learning compared to having no quizzes at all. The second aim is to investigate the role of quiz frequency (weekly, biweekly and monthly) to establish whether more frequent quizzes lead to more vocabulary learning. The study takes into account key individual differences: motivation and self-regulation. It also examines the effect of study time and whether more study time leads to more vocabulary learning. Finally, it investigates students' attitudes towards frequent quizzes and vocabulary learning from a digital flashcard learning app and whether their attitudes play a role in their vocabulary learning.

Due to gender segregation laws in Saudi public schools which prevent males from accessing female schools, both studies were conducted with male students only. Although the limited research on the effect of gender on vocabulary knowledge shows mixed findings (see section 6.3), the lack of female participants may limit the generalizability of the findings.

The final chapter (Chapter 6) provides a general discussion and conclusion which synthesizes the findings from the two studies and relates them to previous research. It also discusses the implications of the two studies and provides suggestions for future research.

2. General literature review

This literature review chapter provides a review of relevant vocabulary research. As outlined in the introduction, the chapter has three sections: the first focuses on fundamental concepts in vocabulary research. The second addresses how vocabulary learning is measured. The third covers the role of individual differences in vocabulary learning. This chapter functions as a literature review for study 1 and study 2 as it is relevant to both (e.g., both studies take into account the role of individual differences in vocabulary growth). Study 2 has an additional literature review that focuses on topics relevant only to the second study (e.g., the testing effect and digital flashcard learning).

2.1 Fundamental concepts in vocabulary research

2.1.1 What is a word?

Vocabulary includes both single words (e.g., apple and go) and multi-word units such as phrasal verbs (e.g., go on and keep up) and idioms (e.g., piece of cake). These multi-word units are often called formulaic language and they are central for fluent second language use (Conklin & Schmitt, 2012; Siyanova-Chanturia & Pellicer-Sánchez, 2018; Wood, 2015). The focus of the thesis however will be on learning, teaching and researching of single words.

Although the concept of a word seems straightforward to understand at first, defining what constitutes a word systematically is challenging (Nation & Webb, 2011; Schmitt, 2010).

Even if we focus exclusively on written language and attempt to define words based on, for example, the blank spaces around them in writing we would immediately run into issues with cases like *well-being* and *long-term* which do not fit neatly to this definition. Another issue is whether to treat the two instances of a word like *care* (noun) and *care* (verb) as one or two words. The difficulties in defining and counting words have led vocabulary researchers to suggest different units of counting to deal with these challenges. The narrowest unit of counting is word-form which treats every instance/form of a word as a separate word. For example, the verbs *care* and *caring* are treated as different words. Next is a lemma, which treats words with different inflections as one word, so the verbs *care* and *caring* are treated as one word but *care* the noun is treated as a different word. A flemma groups inflections and similar spelling derivations together so *care* (n), *care* (v) and *caring* are treated as one word. The broadest unit is a word-family, which groups all the inflections and derivations together as one word. The word-family of the word *care* thus includes *care* (n), *care* (v), *caring*, *careful*, *uncareful*, *careless*, *carelessness* and *uncaring*.

The choice of unit of counting has important implications for vocabulary instruction (e.g., which other forms of a word to focus on during instruction), assessment (e.g., vocabulary size estimates would increase or decrease depending on the unit of counting) and research (Dang, 2021; Stoeckel, Ishii, et al., 2020; Webb, 2021c). Thus, it is no wonder that the topic has been extensively discussed and debated (Dang, 2021; McLean, 2018; Nation, 2021; Stoeckel, Ishii, et al., 2020; Stoeckel, McLean, et al., 2020; Webb, 2021c, 2021a, 2021d). One group of researchers argues that word-families are an appropriate unit of counting (Laufer, 2021; Nation, 2021; Webb, 2021d) while another group of researchers

argues for the more conservative units of counting (i.e., lemma or flemma; Brown et al., 2022; McLean, 2018). The main issue behind the debate according to Nation (2021) is learners' knowledge. Word-family proponents suggest that if a learner knows the headword care then they would not face great difficulties understanding the meaning of the other forms (e.g., caring or careless). On the other hand, the lemma/flemma proponents argue that such an assumption is not valid empirically, and knowledge of one form of a word (e.g., care) does not necessarily entail knowledge of other forms (e.g., uncaring; McLean, 2018). A middle ground has been proposed where the choice of the unit of counting should depend on factors such as the proficiency of the learners and the purpose of the research (Kremmel, 2021; Nation, 2021; Webb, 2021d). For example, word-family might be more appropriate with more advanced learners given that they tend to have a more developed morphological knowledge, while flemma or lemma might be more appropriate for beginners whose morphological knowledge is still limited. Although the recent commentaries and discussion have presented valuable insights and perspectives on the issue of the unit of counting, most of the researchers involved in this discussion agree that there is a need for more empirical studies to help researchers and instructors make more informed decisions about the unit of counting for teaching and research (Kremmel, 2021; Laufer, 2021; Webb, 2021a).

2.1.2 Some words are more useful than others

The English language has around 54,000 word-families (Goulden et al., 1990), which is a very large number that native speakers fall short of reaching (a 20-year-old native speaker is estimated to know approximately 11,100 word-families; Brysbaert et al., 2016). Fortunately, learners do not need to learn all of these words to use language. This is because

only a very small proportion of these words occur very frequently in the language while the majority appear infrequently. High frequency vocabulary encompasses the most frequent 3000 word-families in corpora such as the British National Corpus frequency wordlist developed by Nation (2006). The words between the most frequent 3000 and 9000 word-families are referred to as mid-frequency vocabulary (Schmitt & Schmitt, 2014). The words beyond the most frequent 9000 word-families are termed low frequency vocabulary. The high frequency vocabulary is considered the most useful because of lexical coverage or "how much unknown vocabulary can be tolerated in a text before it interferes with comprehension?" (Nation, 2006, p. 61). It covers up to 85% of most spoken and written language (Adolphs & Schmitt, 2003; Nation, 2006; Webb & Rodgers, 2009a). Although 85% might seem large, research suggests that knowledge of 95-98% of the words in a text is a necessary component of written and spoken language comprehension (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt et al., 2017; Webb & Rodgers, 2009a). Mid-frequency vocabulary is thus considered the second most important vocabulary to learn for general language proficiency because it helps learners reach the 95-98% comprehension threshold (Schmitt & Schmitt, 2014).

The order in which high frequency vocabulary is taught is important. This is because the first 1000 frequency band has much more coverage (77.96%) than the second 1000 frequency band (8.01%), as shown in Table 1, which in turn has more coverage than the third 1000 frequency band (4.36%; Nation, 2022). For mid and low frequency bands, the order is not very important since there are no substantial differences in coverage (e.g., the seventh band coverage is 0.45% and the eighth band coverage is 0.33%).

Table 1. Coverage of the British National Corpus attained by the most frequent 9000 frequency bands with word-families as a unit of counting.

Frequency band	Percentage of coverage of tokens	Percentage of cumulative tokens coverage
1000	77.96	81.14
2000	8.01	89.24
3000	4.36	93.06
4000	1.77	95.37
5000	1.04	96.41
6000	0.67	97.08
7000	0.45	97.53
8000	0.33	97.86
9000	0.22	98.08

Note. Based on Nation (2022).

Even though learning 3000 word-families is more manageable than learning 20,000 word-families, the number is still large for beginners. A more reasonable target in the early stages of language learning is learning the Essential Word List (EWL; Dang & Webb, 2016), which consists of a smaller number of high frequency vocabulary. The EWL is useful because it has fewer words (800 lemmas) and provides a 75% coverage of English discourse. The cutoff point for number of the words in the list was based on three criteria: change in lexical coverage curve (i.e., where including more words does not lead to substantially more coverage), amount of vocabulary needed to understand spoken and written language and practicality. The EWL further breaks down the 800 lemmas into 50 lemma sub-lists to make it easier for teachers to distribute the learning of the items

throughout a language learning course. The items in the list include 176 function words (e.g., the, you, in, at, below, mine, nine, since) and 624 lexical words (e.g., verbs: know, go, get; nouns: thought, life, environment; adjectives: good, special, western; adverbs: well, very, nearly).

Although frequency is useful in guiding which words give the best return for the time invested, it should not be the only determining factor. Some words are useful for the classroom environment despite being of low frequency (e.g., pencil, blackboard; Schmitt & Schmitt, 2020), hence should not be discarded just because they are not frequent (see knowledge-based wordlist; Schmitt et al., 2021). Additionally, some learners might benefit from specialized wordlists that focus on the most frequent words in a specific area. For example, students pursuing higher education might benefit from The Academic Word List (Coxhead, 2000), which includes common academic words that are not specific to any particular field (e.g., approach, deduce, minimal). These words cover approximately 10% of the words in academic texts which translates to roughly one in every ten words, making learning this vocabulary worthwhile (Nation, 2022). Besides academic vocabulary, there is technical or specialized vocabulary which comprises words that are common in a specialist field of study, work or any other community of practice. These may include words that are common in fields like law, aviation, engineering or the military but which are likely to be infrequent in general language use, or may be everyday words used in a technical sense (Nation, 2022). Learning these words can help learners become familiar with the common vocabulary that they are likely to encounter in a specific area of interest.

In sum, not all words are of the same value for all learners and some words (high-frequency, academic and technical) are more useful than others to the level that justifies devoting more time to learning them.

2.1.3 Defining vocabulary knowledge

A clear understanding of what vocabulary knowledge involves is one of the main objectives of vocabulary research. Gass and Selinker point out that "[t]he major task of second language lexical research is to discover what second language learners know about the lexicon of the second language, how they learn it, and why this particular path of development is followed" (2008, p. 545). Vocabulary knowledge has been conceptualized in different ways (Anderson & Freebody, 1981; Daller et al., 2007; Henriksen, 1999; Melka, 1997; Nation, 1990; Richards, 1976). These different conceptualizations should not be viewed as being mutually exclusive but as providing multiple perspectives on the concept of vocabulary knowledge (Durrant et al., 2022).

A common conceptualization is one where vocabulary knowledge is divided into receptive and productive knowledge (Melka, 1997). Receptive vocabulary refers to the collection of words that a learner can understand in listening or reading. Productive vocabulary refers to the collection of words that a learner can use in speaking or writing. Schmitt (2020) suggests that this distinction has ecological validity in that teachers and learners are usually aware of the phenomenon of being able to understand a word but not being able to use it in speaking or writing. Some studies have provided evidence for this observation and shown that learning vocabulary receptively is usually easier than learning vocabulary productively (Laufer & Goldstein, 2004). Also, learners' receptive vocabulary tends to be

larger than their productive vocabulary (Laufer, 1998; Laufer & Paribakht, 1998; Waring, 1997; Webb, 2005). Exactly how large the difference is unknown, but estimates vary from as low as 16% of the receptive vocabulary being known productively to as high as 92% (Schmitt & Schmitt, 2020). The variation in estimates is primarily due to differences in how the two concepts are operationalized and measured (Schmitt & Schmitt, 2020).

Another widely known conceptualization of vocabulary knowledge is distinguishing between vocabulary breadth and depth (Anderson & Freebody, 1981). Vocabulary breadth refers to the number of words known by learners which is commonly referred to as vocabulary size. Vocabulary depth on the other hand refers to how well learners know these words. Schmitt (2014) reviewed studies that investigated the development of vocabulary breadth and depth. His review showed that there is usually no developmental gap between vocabulary breadth and depth for high frequency words (i.e., the more frequent a word, the more deeply it is known). However, for lower frequency words, vocabulary depth appears to lag behind vocabulary breadth growth. The gap is possibly due to low amounts of exposure for lower frequency words compared to higher frequency words.

The concept of vocabulary depth has been examined further to uncover its internal structure (Henriksen, 1999; Nation, 1990, 2022; Richards, 1976). Two main approaches have been followed (Read, 2000): a developmental approach where vocabulary depth is conceptualized as ranging from no knowledge of a word to full knowledge, and a componential approach where vocabulary depth is conceptualized as being formed of smaller components (Nation, 1990, 2022; Richards, 1976). Following the componential approach, Nation (1990, 2022) created the most common and widely used classification of

vocabulary depth knowledge. Nation's framework divides vocabulary knowledge into three main categories, with each further divided into three subcategories:

- Form
    - Spoken
    - Written
    - Word parts
- Meaning
    - Form and meaning
    - Concept and referents
    - Associations
- Use
    - Grammatical functions
    - Collocations
    - Constraints

Form is concerned with knowing how a word sounds (pronunciation) and looks (orthography) as well as knowing its derivations and inflections (morphology). Meaning involves knowing the form and meaning connection (being able to link a form to a meaning), the concepts and referents (what object or objects in the world a word refers to and what concepts are included) and associations (e.g., knowing the relationship between a word and the other words in the language such as synonyms and antonyms). Use involves knowing the grammatical functions (e.g., grammatical category and transitivity for verbs), collocations (what other words frequently co-occur with a word) and constraints (e.g.,

register constraints). Vocabulary knowledge is therefore complex and one of the assessment issues that will be addressed later in section 2.2 is the fact that no test is capable of measuring all of the aspects and nuances of vocabulary. Therefore, when it comes to assessment, Read (2000) points out that multiple instruments need to be used to meet the requirements of vocabulary assessment.

Although useful, this framework is not without limitations. One of these limitations is the lack of a clear specification for the relationships between the different vocabulary knowledge aspects (Schmitt, 2019) which hinders a fuller understanding of vocabulary knowledge. Also, the framework does not specify the order of the acquisition among the aspects (Schmitt, 2019). This limits its pedagogical value and makes it less informative in terms of choosing which vocabulary aspects should be prioritized when teaching. Despite these limitations, the framework remains useful in guiding vocabulary instruction and research given that it breaks vocabulary knowledge into more manageable components.

2.1.4 Vocabulary and language proficiency

It is widely accepted that vocabulary is essential to all language use (Schmitt et al., 2017) including reading (Laufer & Aviad-Levitzky, 2017; Laufer & Ravenhorst-Kalovski, 2010; D. Qian, 1999; D. D. Qian, 2002; Stæhr, 2008), listening (Bonk, 2000; Y. Li & Zhang, 2019; Stæhr, 2008; van Zeeland & Schmitt, 2013; S. Zhang & Zhang, 2022), viewing TV (Durbahn et al., 2020; Peters & Webb, 2018; Teng, 2022), writing (Laufer, 1994; Stæhr, 2008) and speaking (De Jong et al., 2012; Uchihara & Clenton, 2023; Uchihara & Saito, 2019). These studies have had the common goal of understanding the relationship between vocabulary knowledge and language skills, but they have approached this goal differently.

Some studies have focused on vocabulary breadth as a measure of vocabulary knowledge (Miralpeix & Muñoz, 2018; Stæhr, 2008) while others have examined both simultaneously (C. Chen & Liu, 2020; M. Li & Kirby, 2015; D. Qian, 1999; D. D. Qian & Schedl, 2004; Wu et al., 2021). The discussion here will focus on vocabulary breadth or size since it is the focus of the thesis (for research on vocabulary depth, see Dóczi & Kormos, 2015; M. Li & Kirby, 2015; Milton & Fitzpatrick, 2014; Schmitt, 2014).

The previous studies suggest that having a large vocabulary is important for all language skills. How large the vocabulary needs to be for 'adequate'[1] comprehension and production varies depending on, for example, the skill investigated (listing, reading, writing or speaking), unit of counting (e.g., lemma or word family), genre (e.g., academic, literary), material level (simplified or unsimplified) and incompatible researcher findings (estimates sometimes differ from one researcher to another; Laufer & Ravenhorst-Kalovski, 2010; Webb, 2021b).

One of the key factors here is text coverage (see section 2.1.2). To reach 98% coverage of texts, Nation (2006) suggests that 8000 to 9000 word families are needed for unassisted reading comprehension and 6000 to 7000 for listening. To understand spoken discourse and everyday conversations, a vocabulary size between 2000 to 3000 word-families is needed (95% coverage; van Zeeland & Schmitt, 2013). For writing, the first 2000 words from the General Service List[2] (GSL; West, 1953) were found to account for the majority of English language learners' writing (Laufer, 1994; Laufer & Paribakht, 1998). As these

[1]What is 'adequate' can vary depending on several factors such as coverage level and text genre. See Laufer & Ravenhorst-Kalovski, (2010)

[2] A list of the most frequent words in English comprising around 2000 words.

learners become more proficient, the proportion of words beyond the GSL starts to increase. Estimating the vocabulary size needed for speaking (and perhaps writing) is more challenging due to the fact that during speaking learners have control over what language they choose and can use certain strategies when their vocabulary is limited such as circumlocution (i.e., using multiple words to explain a concept when the single word is unknown) and gestures (Schmitt & Schmitt, 2020). Therefore, the amount of vocabulary needed for speaking will mostly depend on factors such as the topic and the learner's use of strategies. Schmitt (2020) suggests that the best that can be done until further research is conducted is to assume that the vocabulary size needed for the receptive skills is sufficient for the productive ones. Based on the previous discussion, it can be seen that a sizeable vocabulary is essential for performance in the four skills of language.

2.1.5 Incidental and intentional vocabulary learning

There are two main approaches to vocabulary learning: incidental vocabulary learning and intentional vocabulary learning (Nation, 2022). Incidental learning refers to learning words as a by-product of a task as when reading a book or watching a movie (Ellis, 1999; Hulstijn, 2003) whereas intentional learning occurs when the goal of a task is to learn language features such as vocabulary (e.g., learning vocabulary from wordlists). The primary sources of incidental vocabulary learning include reading (Al-Homoud & Schmitt, 2009; Pellicer-Sánchez, 2016; Swanborn & De Glopper, 1999; Webb, 2005), listening (Bonk, 2000; Brown et al., 2008; Elley, 1989; Pavia et al., 2019; Smidt & Hegelheimer, 2004) and watching television (Montero Perez et al., 2018; Peters & Webb, 2018; Puimège & Peters, 2019; Teng, 2022; A. Wang & Pellicer-Sánchez, 2022). Common sources of intentional

vocabulary learning include learning from flashcards (McLean et al., 2013; Nakata, 2008) and wordlists (Nakata, 2008; Webb, 2007a, 2009)

Native speakers learn the majority of their vocabulary through incidental learning (Nagy et al., 1987; Sternberg, 1987). They add roughly 1000 word-families a year between the ages of three and 15 (Goulden et al., 1990; Nation, 2022) or approximately 1.7 new word-families a day after the age of two (Brysbaert et al., 2016), which is not surprising given the large amount of language input available to them. EFL learners on the other hand lack access to widespread input, which has led some to argue that they learn the majority of their vocabulary through intentional vocabulary learning (Laufer, 2003, 2005). Given that the time allocated for foreign language classes is usually limited (Lightbown & Spada, 2020), foreign language learners end up learning substantially less vocabulary than native speakers (see section 2.2.4).

Within vocabulary research, there is generally a consensus that L2 vocabulary should be learned both incidentally and intentionally (Barclay & Schmitt, 2019; Nation, 2011; Siyanova-Chanturia & Webb, 2016; Webb & Nation, 2017). This is because each approach has its own strengths and limitations. Intentional vocabulary learning can be very efficient in the sense that many words can be learned in a relatively short amount of time (McLean et al., 2013; Nakata, 2008, 2020). For example, McLean et al. (2013) showed that learners were able to learn 1107 word-families (similar to annual gains of native speakers) from flashcards over an academic year (see section 5.1.1 for more discussion on flashcard learning). Especially for high frequency vocabulary, intentional vocabulary learning can be an effective approach to boost students' vocabulary knowledge. Despite being efficient, intentional vocabulary learning is time-consuming and it would be impractical to teach all

words learners need to master as well as all aspects of vocabulary knowledge such as collocations and constraints intentionally given the limited time available for foreign language instruction (Schmitt & Schmitt, 2020). Incidental vocabulary learning (e.g., from reading), on the other hand, is more context-rich, where not only is there an opportunity for form-meaning link knowledge to develop (Pellicer-Sánchez, 2016), but also other aspects of vocabulary knowledge such as collocation and register (e.g., knowing that a word is more common in formal contexts than in everyday language; Webb et al., 2013). One main limitation of incidental vocabulary learning is that the vocabulary gains made from this approach are low compared to intentional vocabulary learning (Nation, 2022; Schmitt & Schmitt, 2020; Webb et al., 2023). Therefore, combining both incidental and intentional learning in a vocabulary program ensures a more balanced approach where students can benefit from the efficiency of intentional vocabulary learning and from the context-rich learning of incidental vocabulary learning (Nation, 2022).

2.2 Measuring vocabulary knowledge

Vocabulary testing is important for vocabulary research, which is evident in the fact that vocabulary tests are among the most used research instruments (Durrant et al., 2022). Whole books (Milton, 2009; Read, 2000), book sections (Durrant et al., 2022; Nation, 2022; Schmitt & Schmitt, 2020; Webb, 2019) and several articles (e.g., Read & Chapelle, 2001; Schmitt et al., 2019; Stoeckel, McLean, et al., 2020) have been written on the topic to help move the field of vocabulary testing forward. One of the main outcomes of vocabulary testing research is the development of more accurate vocabulary tests (for a more detailed review see: Durrant et al., 2022; Milton, 2009; Nation, 2022; Schmitt, 2020; see also section 3.2 below).

Given the difficulty in measuring all aspects of vocabulary knowledge at once (Read, 2000), test developers tend to focus on one or a few aspects of vocabulary knowledge when designing tests. The currently available researcher-developed vocabulary tests have followed the conceptualizations and distinctions of vocabulary knowledge discussed in section 2.1.3 which can be used to classify them. The first of these is the distinction between tests that measure vocabulary breadth and tests that measure vocabulary depth. Tests that focus on breadth give estimates of how many words learners know by measuring the form-meaning component (e.g., form is provided and learners supply the meaning). Tests focusing on depth tell us how well these words are known by measuring the other vocabulary knowledge components (e.g., asking learners not only to provide a word meaning, but also other components such as its collocations or associations). The second distinction is between receptive and productive vocabulary knowledge tests. Receptive knowledge tests assess learners' ability to recognize the meaning of a word in reading or listening while productive knowledge tests test learners' ability to use a word in speaking or writing (Nation, 2022; Schmitt, 2010). The final distinction is between tests that measure meaning recognition and meaning recall knowledge of words. Meaning recognition tests require learners to choose the correct form or meaning of a word, while meaning recall tests require learners to retrieve from memory a word meaning or form. Focusing on vocabulary breadth, the combination of receptive/productive and recognition/recall knowledge mastery in Table 2 provides an overview of the four possibilities of vocabulary knowledge tests of form-meaning (Laufer & Goldstein, 2004; Schmitt & Schmitt, 2020).

Table 2. Vocabulary knowledge of form-meaning test types

	Receptive Recognition	Productive Recognition	Receptive Recall	Productive Recall
Provided	Form	Meaning	Form	Meaning
Tested	Meaning recognition	Form recognition	Meaning recall	Form recall
Example	1- Car a- furniture b- vehicle c- container	1- A type of vehicle a- car b- chair c- spoon	1- license ……..	1- a permit to use or own something l…….

Before exploring the vocabulary knowledge tests available, it is important to review some key considerations in vocabulary assessment. This is important because some vocabulary tests have certain design issues and some have not been properly validated before publication (Durrant et al., 2022; Schmitt et al., 2019). Schmitt notes that "most vocabulary tests are not validated to any great degree" (2019, p. 268). The next sections aim to provide a critical evaluation of vocabulary tests and discuss their strengths and weaknesses.

2.2.1 Key issues and considerations in vocabulary assessment

Vocabulary tests, like any other language test, need to meet three key criteria before they can be used: validity, reliability and practicality (Bachman, 1990; Bachman & Palmer, 1996; Read, 2000). Validity is a multifaceted and complex construct (Messick, 1989), one key condition of which is that a test needs to measure what it is supposed to measure and minimize influence from irrelevant factors (Milton, 2009; Schmitt, 2010). For example, if the intended construct of measurement is productive vocabulary knowledge, then receptive knowledge should not interfere significantly in the measurement (e.g., free recall reflects more accurately productive knowledge than cued recall). Reliability refers to consistency and stability in measurement. In other words, a test should give similar results if for example it was taken multiple times in the same session by the same learner. Similarly, if a test has two versions, they need to give similar results if they were taken by the same learner on the same day. Finally, practicality refers to the condition of efficiency in that a test for example should not take too much resource to administer and score. This is why, for example, most vocabulary tests focus on one aspect of vocabulary knowledge because attempting to test all aspects reliably would probably take too much time to administer. The three criteria of validity, reliability and practicality should be taken into consideration when evaluating the different vocabulary tests available. In addition to the broader language testing considerations, there are more vocabulary-focused issues that need to be considered, including the number of vocabulary aspects to test, item format, sampling rate and the influence of cognates (Durrant et al., 2022; Stoeckel, McLean, et al., 2020).

2.2.1.1 Aspects of vocabulary knowledge

Most vocabulary knowledge tests focus on breadth of vocabulary knowledge and form-meaning link, which has been criticized by some (e.g., Milton & Fitzpatrick, 2014) on the grounds that it does not capture the full complexity of vocabulary knowledge. Despite this drawback, the choice of measuring breadth of vocabulary and form-meaning knowledge is not unjustifiable. First, focusing on one aspect means that more items can be tested, and the test can be more representative of this aspect (Read, 2000). Second, the form-meaning link is the most important aspect of vocabulary knowledge given that meaning errors (e.g., referring to a cat as a dog) are usually more severe in terms of comprehension than grammatical errors (e.g., using the wrong word form: *what is the different between x and y; Laufer & Goldstein, 2004).

2.2.1.2 Item format

Most vocabulary breadth tests measure vocabulary knowledge using meaning recognition tests which has also been criticized on the grounds that they tend to overestimate the number of words learned due to random guessing (Gyllstad et al., 2015; Stoeckel, McLean, et al., 2020). However, the opposite might happen if vocabulary knowledge is measured using only recall tests since they tend to underestimate the number of words learned (Kremmel & Schmitt, 2016). A more serious issue with meaning recognition tests relates to ecological validity in that they might not reflect receptive vocabulary knowledge reliably (Gyllstad et al., 2015; Kremmel & Schmitt, 2016; Stewart, 2014). When learners use language receptively (reading or listening), they are not offered a list of meaning choices to choose from, but they must recall word meaning from the mental lexicon. How

representative meaning recognition tests are of receptive vocabulary knowledge remains open for further research (Stewart et al., 2021; Stoeckel, McLean, et al., 2020; Webb, 2021a).

2.2.1.3 Sampling rate

Given the impracticality of testing all words in frequency bands in a vocabulary test (which tend to be in the thousands), test developers normally resort to sampling 10 to 40 words from each frequency band and test these words only. When learners answer most of these words correctly, it is assumed that they know the majority of the other words in the frequency band. These assumptions are based on the finding that learners tend to learn more frequent words (e.g., house) before learning less frequent words (e.g., dwelling). Based on this, it has been hypothesized that if learners know a word in one frequency band (e.g., expensive from the first 1000 band) there is a good chance that they know the other words from the same frequency band (e.g., good, happy, hot). A key issue here is sampling rate or how many words should be tested from a frequency band to be deemed representative of mastery of the majority of words in that frequency band (Gyllstad et al., 2015; Stoeckel, McLean, et al., 2020). Vocabulary tests vary between as little as 5 words per frequency band to as high as 40. One possible recommendation is ‘the more the better’, however, practicality would soon become an issue (Durrant et al., 2022). A more practical and seemingly sufficient threshold is 30 words per frequency band (Gyllstad et al., 2015, 2021). In Gyllstad et al. (2021) 103 Japanese EFL learners were tested on all the words in a frequency band (3000 band) using meaning recall and meaning recognition tests. Using

bootstrapping[3], they compared tests with 5, 10, 20, 30, 40, 50, 60, 100 and 200 items to the students' actual test scores. They found that the mismatch between the bootstrap samples and the actual test scores declines as test items increase. The percentage of difference was highest with a sampling rate of five items (50% for meaning recall test and 20% for meaning recognition) and least with the 200-item test (10% for meaning recall test and 5% for meaning recognition). More importantly, they found that the curve starts to flatten out after the 30-item threshold. Based on this, they recommend a sampling rate of 30 words per frequency band for both meaning recognition and meaning recall vocabulary tests.

2.2.1.4 The effect of L1

When language learners take vocabulary tests, they bring with them their L1 resources which can influence the test scores. Two main areas have been investigated in this regard: the role of translation and cognates (Durrant et al., 2022; Read, 2019). The translation of vocabulary tests to learners' L1 (creating bilingual vocabulary tests) has been supported by Nation (2022) since the 1990s on the grounds that this might minimize the influence from factors other than vocabulary knowledge (e.g., knowledge of relative clauses, see: Nguyen & Nation, 2011) which should enhance the construct validity of the test. Following Nation's recommendation, a number of bilingual vocabulary tests have been developed for several languages such as Vietnamese (Nguyen & Nation, 2011) and Persian (Karami, 2012). Elgort (2013) provided evidence for Nation's recommendation when she compared a monolingual vocabulary test with a Russian bilingual vocabulary test. 121 intermediate

[3] Bootstrapping is "a type of robust statistic that simulates how a study would be replicated by resampling from a population." (LaFlair et al., 2015, p. 46)

EFL learners took both tests (70 items in each) and their results showed significantly higher scores (32.97) on the bilingual test than the monolingual one (29.61). Her findings suggest that giving a monolingual test can significantly underestimate the vocabulary knowledge of learners by up to 672 word-families. Thus, bilingual vocabulary tests might be more sensitive measures of vocabulary knowledge than monolingual tests.

The second area where the role of L1 was examined is the effect of cognates on vocabulary test scores. Cognates or loanwords are words that share a similar sound and meaning in two languages (Laufer & McLean, 2016). For example, the Spanish word persona and the English word person are considered cognates because they have a similar phonological form and meaning across the two languages. The two terms cognates and loanwords are often used interchangeably. However, when talking about two genetically unrelated languages such as Arabic and English, the term loanwords might be more appropriate since these languages do not share a common ancestor (Laufer & McLean, 2016). One of the most common areas where other languages have borrowed words from English is in the area of technology. For example, the Arabic words *televizion* (television), *fedio* (video) and *combuter* (computer) are loanwords that were borrowed from English.

When it comes to vocabulary testing, cognates and loanwords pose a challenge for test developers and researchers. Cognates and loanwords tend to be answered more correctly than non-cognates and non-loanwords (Allen, 2018, 2019a, 2020; Elgort, 2013; Laufer & McLean, 2016). In itself this is not an issue given that cognates and loanwords are part of the learners' lexicon and they should be represented in the vocabulary knowledge estimates (Nation & Webb, 2011). However, it might become a problem when the proportion of cognates and loanwords in a test is not representative of their proportion in the language

(Cobb, 2000; Laufer & Levitzky-Aviad, 2018). This can lead to either overestimation or underestimation of vocabulary knowledge. For example, Elgort (2013) found that the proportion of English-Russian cognates in a vocabulary test was 34% which is higher than the 27% proportion found in the wordlist which the test items were sampled from. This can lead to overestimation in vocabulary knowledge (Allen, 2019a; Elgort, 2013). One solution that has been followed is to develop a customized vocabulary test for a homogenous group of EFL learners who share a common L1 which takes into account the accurate proportion of cognates (Peters et al., 2019). The situation is more complicated when a group of learners have different L1s (Laufer & McLean, 2016), and no solution appears to be viable that ensures an accurate representation of cognates in the vocabulary knowledge estimates in this case.

No research seems to have been conducted on the effect of loanwords on the English vocabulary test scores of Arabic-speaking learners. Nevertheless, research on Hebrew speakers might provide some useful estimations of the loanwords effect given that both Arabic and Hebrew are genetically related languages and belong to the same Semitic language family. Laufer and Levitzky-Aviad (2018) examined how the presence of English-Hebrew loanwords affected the vocabulary test scores of 303 Hebrew EFL learners with three levels of proficiency. The learners took tests with varying numbers of loanwords, including tests with no loanwords, tests with a representative number of loanwords and tests with a random number of loanwords. These tests covered four aspects of vocabulary knowledge: form recall, meaning recall, form recognition, and meaning recognition. The results showed that the impact of loanwords on test scores varied depending on the specific modality of the test and the proficiency levels of the learners. The key finding is that the

score increase from the version of the test with representative loanwords to the version with random loanwords was minimal, and the differences in the effect size were very small. Therefore, overall, the study suggests that loanwords in vocabulary tests may not significantly affect the accuracy of measuring true vocabulary knowledge. Overall, although cognates and loanwords have a significant facilitating effect that tends to inflate English vocabulary test scores, the magnitude of the effect seems to depend on the L1 of the learners. The influence appears to be minimal for some languages such as Hebrew (Laufer & Levitzky-Aviad, 2018) and larger for genetically related languages such as French (Cobb, 2000) and languages with more borrowings from English such as Japanese (Allen, 2019a, 2019b; Daulton, 2008).

In summary, like any language test, vocabulary tests need to meet the criteria of validity, reliability and practicality. In addition to these criteria, vocabulary teachers and researchers need to be aware of other factors that might have an influence on vocabulary testing such as item format, sampling rate, translation and cognates. Having reviewed these key concepts and issues, the vocabulary tests discussed in the next section can be better evaluated and critically examined.

2.2.2 Common standardized receptive vocabulary breadth tests

Since the 1980s, several vocabulary breadth tests have been developed. The following list shows some of the commonly used tests of vocabulary form-meaning knowledge (see Figure 1 for sample tests items):

- Vocabulary Levels Test (Nation, 1983, 1990; Schmitt et al., 2001; Webb et al., 2017)
- Checklist tests (Meara, 1992; Meara & Jones, 1988)
- Computer Adaptive Test of Size & Strength (Aviad-Levitzky et al., 2019; Laufer & Goldstein, 2004)
- Vocabulary Size Test (Nation & Beglar, 2007)

Figure 1. Sample items from common vocabulary breadth tests

Test — **Item sample**

VLT

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

VST

1. innocuous: This is **innocuous**.
   - a cheap and poor in quality
   - b harmless
   - c not believable
   - d very attractive-looking

CATSS

- Recall of word form: She is a ***l***_____ girl. (small)
- Recall of word meaning: 'She is a ***little*** girl' means that she is _____
- Recognition of word form: She is a _____girl. (small)

a. little b. great c. nice d. single

- Recognition of word meaning: 'She is a ***little*** girl' means that she is _____

a. small b. great c. nice d. single

Checklist test

What you have to do:

Read through the list of words carefully. For each word:

if you know what it means, write Y (for YES) in the box

if you don't know what it means, or if you aren't sure, write N (for NO) in the box.

1 ❑ high	**2** ❑ building	**3** ❑ possible
4 ❑ fear	**5** ❑ rope	**6** ❑ attard
7 ❑ nice	**8** ❑ neighbour	**9** ❑ general
10 ❑ lazy	**11** ❑ equalic	**12** ❑ cordle

VLT (Webb et al., 2017); VST (Nation & Beglar, 2007); CATSS (Aviad-Levitzky et al., 2019); Checklist tests (Meara, 1992)

A key distinction in vocabulary breadth tests is made between vocabulary levels tests and vocabulary size tests (Milton, 2009; Nation, 2022). Vocabulary levels tests (Nation, 1983, 1990; Schmitt et al., 2001; Webb et al., 2017) were developed for diagnostic purposes in

that they provide information regarding which frequency levels or specific words (e.g., academic words) students know well and which need further development. On the other hand, vocabulary size tests (Aviad-Levitzky et al., 2019; Meara, 1992; Nation & Beglar, 2007) were developed to provide information about learners total vocabulary knowledge.

Earlier tests (Laufer & Goldstein, 2004; Nation, 1983, 1990; Schmitt et al., 2001) relied on wordlists that were based on small (nowadays considered outdated) corpora such as the GSL to determine word frequency. With the advent of computerized and large corpora such as the British National Corpus and the Corpus of Contemporary American English, more accurate and up-to-date wordlists were created (Nation, 2006) which later tests (Aviad-Levitzky et al., 2019; Nation & Beglar, 2007; Webb et al., 2017) relied on.

The Vocabulary Levels Test (VLT; Nation, 1983, 1990; Schmitt et al., 2001) is possibly the most widely used test of learners' vocabulary knowledge (Read, 2000). The early versions measure learners' receptive knowledge of words at four frequency levels (2000, 3000, 5000 and 10,000) and their knowledge of academic words. The Updated Vocabulary Levels Test (UVLT; Webb et al., 2017) differs from previous VLTs in that it uses updated wordlists and measures every 1000 frequency level from the first 5000 words (previous VLTs skipped the first and fourth frequency levels). This comes at the expense of excluding the 10,000 frequency level and the academic vocabulary part. One of the test's strengths lies in that it has a higher sampling rate (30 items per 1000 frequency level) compared to other tests, which can provide more accurate vocabulary knowledge estimates (Gyllstad et al., 2015, 2021). Two versions from the UVLT were initially validated by Webb et al (2017) on 250 university students from three countries (China, Japan and Spain). The results suggest that the test is a valid (e.g., the test difficulty increases as words' frequency

decreases) and reliable measure of written receptive knowledge of form-meaning link (Rasch subject and item reliability = .96).

Checklist tests (Meara, 1992; Meara & Jones, 1988) measure learners' vocabulary size by presenting sample words from different frequency levels and asking learners to check the ones they know. These tests differ from other tests in that learners are not required to demonstrate their knowledge of words, which can be problematic since some learners might overestimate their lexical knowledge. One common convention to overcome this shortcoming is by including nonwords (i.e., made-up words) to adjust the overall score for possible overestimation. Some checklist tests, such as X-Lex (Meara & Fitzpatrick, 2000), have no validation records.

Most vocabulary tests have the limitation of testing only one aspect of vocabulary knowledge (e.g., receptive recognition). However, The Computer Adaptive Test of Size & Strength (CATSS; Aviad-Levitzky et al., 2019; Laufer & Goldstein, 2004) overcomes this shortcoming by testing all four aspects of the form-meaning component of vocabulary knowledge. It tests the strength of the form-meaning connection on the basis of four modalities: receptive[4] recognition, productive recognition, receptive recall and productive recall (listed in Table 3 above). The new version of the CATSS uses more updated word lists and measures learners' vocabulary size by sampling 140 words from the first 14 1000 frequency bands. It was validated on 453 university students and appears to be valid and reliable (overall test Cronbach's Alpha = .98). The test's advantage of testing words across four modalities comes at a cost since it suffers from a low sampling rate of only 10 items

[4] Laufer & Goldstein use the term passive for receptive and active for productive

per 1000 band, which some consider insufficient to represent the whole frequency level (Gyllstad et al., 2015; Schmitt et al., 2001). Moreover, the low sampling rate means that this test is less ideal for testing vocabulary growth longitudinally, since newly learned words are more likely to be missed in the items chosen for testing compared to tests with higher sampling rate (e.g., UVLT; Webb et al., 2017).

The Vocabulary Size Test (VST; Nation & Beglar, 2007) measures learners' overall vocabulary knowledge by sampling 10 words from the 1st 1000 frequency band up until the 14th with a total of 140 items. It is a multiple-choice test that measures in particular the written receptive knowledge of the form-meaning link; thus, it does not provide information about the other vocabulary knowledge components. The VST has been suggested to be a reliable and accurate measure of learners' receptive knowledge of the most frequent 14,000 words (Beglar, 2010). However, the same low sampling issue and its implications in the CATTS also apply to the VST.

What the previous discussion shows is that each vocabulary test has advantages and limitations and that there is no single test that will fit all contexts. The choice of the most appropriate test for a specific context depends on several factors including the purpose of use (e.g., to test knowledge of high frequency words, diagnose knowledge of academic words, measure vocabulary size), vocabulary knowledge aspect to test (i.e., receptive recognition, receptive recall, productive recognition or productive recall), learners' proficiency (i.e., tests of high frequency vocabulary might be enough for beginners but mid-frequency is needed with more advanced learners) and time available (e.g., only one vocabulary knowledge aspect such as meaning recall or meaning recognition if time is

limited but perhaps more if time allows it). These factors are likely to play a role when deciding which vocabulary test to use.

2.2.2.1 Key considerations when measuring vocabulary growth

A key question is how what has been discussed so far relates to the measuring of vocabulary growth (which is the focus of study 1, Chapter 4). When tracking vocabulary growth, vocabulary size tests[5] such as the VST and CATTS have the advantage of covering a wide range of frequency bands (1000 – 14,000) which can increase the likelihood of detecting growth. Increasing the frequency bands covered by a test usually comes at the expense of reducing the sampling rate in these tests (e.g., 10 words per 1000 frequency band in the VST and CATTS) which can decrease the likelihood of detecting vocabulary growth. Recent vocabulary levels tests (UVLT and NVLT) have the advantage of a higher sampling rate (UVLT = 30 per 1000 frequency band, NVLT = 24) but have the limitation of less coverage (e.g., both the UVLT and NVLT do not cover words beyond the most frequent 5000 words). To give an example, suppose a language learner in a vocabulary growth study has learned a new word in the 7000 frequency band, this word will not be detected in the vocabulary levels tests since this band is not covered in the tests but it might be detected in vocabulary size tests which cover this band. In terms of the sampling rate, if a learner learned a new word for instance in the 3000 frequency band, this word would be more likely detected in vocabulary levels tests since they tend to have a higher sampling rate than vocabulary size tests.

---

[5] Excluding vocabulary size tests where learners are not required to demonstrate their knowledge of form-meaning link such as checklist tests.

This raises the question of which types of tests (recent levels or size) are more appropriate for research on vocabulary growth. One key factor that might help in making this decision is learners' proficiency. For more advanced learners, vocabulary size tests are more appropriate since learners are more likely to learn mid-frequency (3000-9000) and perhaps low frequency vocabulary (beyond 9000 words). For beginners with a small vocabulary size, sampling rate might be more important than coverage since they are less likely to learn words beyond the most frequent 5000 words. In fact, testing beginners on lower frequency bands might increase random guessing and hence overestimate vocabulary growth (Beglar, 2010; Elgort, 2013; Mclean et al., 2015; Stewart, 2014). Therefore, recent vocabulary levels tests such as the UVLT might be more appropriate for beginners than vocabulary size tests when tracking vocabulary growth.

The term beginners has been used somewhat vaguely here. This is because it is difficult to specify with confidence exactly when vocabulary levels tests are no longer appropriate to be used as vocabulary growth tests. One rule of thumb is to allow for two frequency bands beyond learners' current level (a rule originally suggested to minimize random guessing; Beglar, 2010; Elgort, 2013). Based on this, recent vocabulary levels tests (e.g., UVLT) might be sufficient for learners with a vocabulary size of up to 3000 word-families. It should be noted that as learners' vocabulary size increases, the likelihood of missing newly learned words starts to increase in vocabulary levels tests. Therefore, the UVLT is likely to be sufficient for learners with a vocabulary size of 1000 word-families or lower, but as their vocabulary size increases, the confidence in the growth estimates starts to decrease as they are more likely to learn words beyond 5000 word-families limit.

2.2.3 Vocabulary growth and the importance of longitudinal data

Longitudinal studies on vocabulary acquisition mainly follow the breadth and depth dichotomy discussed before in section 2.1.3. Researchers on vocabulary breadth are primarily interested in knowing how many words are learned over a period of time (Clark & Ishida, 2005; Milton, 2006; Webb & Chang, 2012; X. Zhang & Lu, 2014), while researchers on vocabulary depth are primarily interested in the development of the different components of vocabulary knowledge (Fitzpatrick, 2012; González-Fernández & Schmitt, 2019; Schmitt, 1998; Zheng, 2016).

Longitudinal research can be carried out using both longitudinal and cross-sectional data. Longitudinal data measure research units (in this case learners) over multiple time points while cross-sectional data measure research units at one time point (Taris, 2000). The principal types of longitudinal studies are (Dörnyei, 2007):

- Panel studies (multiple measures at different time periods from the same participants)
- Trend studies (multiple measures at different time periods from different participants)
- Simultaneous cross-sectional studies (single measure from different age groups)
- Retrospective longitudinal studies (single measure based on thinking back technique)

Both longitudinal and cross-sectional data can be used to give information about the development of learners' vocabulary knowledge over time. An example of a cross-sectional study measuring vocabulary breadth growth would involve giving a vocabulary size test to

learners from different levels at the end of a school year or a course period. Results for how many words each level learned are then compared to measure the progress made. For example, if first-year students in a secondary school finished the year with 1000 words and second-year students finished with 1500 words, we might assume that second-year students have learned approximately 500 new words. Cross-sectional data has been used in some longitudinal studies (e.g., Milton, 2006), however, it has two main drawbacks. The first issue relates to the fact that it measures two different cohorts (i.e., in the above example for instance, we do not know for sure that the second-year students started the year with 1000 words, which can compromise the progress estimates). The second issue has to do with the lack of progress estimates for the first group in a data set since it does not have a baseline number to compare to. Longitudinal data (e.g. Webb & Chang, 2012, discussed in the next section), on the other hand, give us more accurate figures since the same learners' vocabulary knowledge is measured at different time intervals. The cohort effect is not normally an issue in panel studies since we are measuring the same population (Taris, 2000). Thus, whenever it is possible, longitudinal data should be used to track vocabulary development (and language overall) since they provide more accurate growth estimates than cross-sectional data.

2.2.4 Research on vocabulary breadth growth

Tracking vocabulary knowledge development over time can provide key insights to research on vocabulary knowledge development. A commonly stated observation within vocabulary research is the fact that the field has yet to develop an overall theory of vocabulary knowledge development (Schmitt, 2019). Therefore, research that tracks how vocabulary develops over time can provide empirical data against which theoretical

assumptions can be tested. Research on vocabulary growth can also have practical implications. For example, it can provide benchmarks for language programs to evaluate the vocabulary growth rate of their learners and whether they are progressing as expected or whether an intervention is needed. Moreover, it can help textbook developers make informed decisions regarding the number of target vocabulary to include in the textbooks (Milton, 2009). Milton and Hopwood (2022) note that many EFL textbooks lack a principled approach to the type and amount of vocabulary to include. For example, studies report that a proportion of important vocabulary (high frequency words) is missing from EFL textbooks with estimates ranging from 30% (in the Success coursebook series; Eldridge & Neufield, 2009) to 15% (Saudi EFL textbooks; Alsaif & Milton, 2012). Despite its importance, this line of inquiry remains relatively under-researched (Dóczi & Kormos, 2015; Pellicer-Sánchez, 2019; Webb & Nation, 2017). Nevertheless, rough vocabulary growth estimates can be made based on the few studies available.

Reviewing the research on vocabulary size growth reveals a number of patterns and limitations (Agustín-Llach & Alonso, 2016; Coxhead & Boutorwick, 2018; Gallego & Llach, 2009; Laufer, 1998; Milton & Meara, 1998; Ozturk, 2012; Robles-García, 2022; Stoeckel, 2018; Webb & Chang, 2012; X. Zhang & Lu, 2014). First, studies have focused mostly on the growth of receptive recognition knowledge (e.g., Webb & Chang, 2012; X. Zhang & Lu, 2014) with few studies examining productive growth (Laufer, 1998), and none appear to have examined other aspects such as meaning recall. Second, while some studies have been conducted with target languages other than English such as French (Milton & Meara, 1998) and Spanish (Robles-García, 2022), the majority of studies had English as a target language. Although this is useful in providing insights into how

vocabulary develops in English, it does not give a comprehensive view of how vocabulary develops in a second language, for which there is a need for research on other languages. Third, most studies were conducted in EFL contexts with relatively few being conducted in English as a Second Language[6] (Robles-García, 2022) and English Medium Instruction contexts (EMI; Agustín-Llach & Alonso, 2016; Coxhead & Boutorwick, 2018). Examining vocabulary growth in different language learning contexts can provide useful insights since it is unlikely that vocabulary growth in an EFL class and an EMI class will be the same (mainly due to differences in the amount of input). Similarly, whilst there has been some commendable diversity in the L1 backgrounds of the participants including Chinese (Webb & Chang, 2012; X. Zhang & Lu, 2014), English (Milton & Meara, 1998; Robles-García, 2022), French (Milton & Meara, 1998), German (Milton & Meara, 1998) Greek (Milton & Meara, 1998), Hebrew (Laufer, 1998) and Japanese (Stoeckel, 2018), learners with other language backgrounds such as Arabic have not been investigated. Research on other language backgrounds is likely to be needed given that L1 has been found to affect L2 vocabulary learning, as in the faciliatory effect of cognates discussed previously (Elgort, 2013; see section 2.2.1). Finally, the most widely-used instrument was older versions of the vocabulary levels test (e.g., Agustín-Llach & Alonso, 2016; Gallego & Llach, 2009; Webb & Chang, 2012; X. Zhang & Lu, 2014), which is less than ideal for vocabulary growth since they skip important frequency bands such as the first 1000 (which covers

[6] ESL is a term used to refer to learners who learn English in a country where it is the native language (e.g. immigrants learning English in the UK). The term EFL on the other hand refers to those who learn English where it is a not the native language of the country (e.g., Saudi students learning English in Saudi Arabia). See (Gass & Selinker, 2008). EMI refers to using English as the medium of instruction in education where English is not the native language of the majority of the population (Saudi students studying medicine or chemistry in Saudi Arabia). See (Smit, 2023).

more than 75% of English use; Nation, 2006; see section 2.2.2) and other important mid-frequency bands (i.e., 4000, 6000, 7000, 8000, 9000). Only Webb and Chang (2012) modified the VLT and included the most frequent 1000 frequency band in the test. Their study and some of the other reviewed studies are discussed further in the following section.

In one of the few longitudinal studies, Webb and Chang (2012) examined the vocabulary breadth development of 166 high school EFL learners over five years in Taiwan using a modified version of the VLT (Schmitt et al., 2001) that included words from the first 1000 band. There were three groups, one majoring in English and two in diverse subjects. Results showed that learners vary in the amount of vocabulary learned over a school year, where two groups of students from diverse subjects performed very low while a group of English students performed relatively higher, learning as many as 430 word-families. Instruction time seems to be a key factor given that the English studies group received an average instruction time of around 15 hours per week while the other two groups received between 2 to 3.4 hours. Overall, after nine years of study, only half of the participants mastered the most frequent 1000 word-families. This percentage goes even lower when moving to the second 1000 words, which only 16% of the participants mastered. Webb and Chang's study remains to date the longest longitudinal study on L2 vocabulary growth. This however came with the cost of being limited in measuring only one aspect of vocabulary knowledge (form-meaning recognition) and the lack of other key exploratory variables such as individual differences or out-of-class vocabulary learning (Pellicer-Sánchez, 2019).

Learners in some countries seem to learn more words than others. For instance, Milton and Meara (1998) compared the vocabulary growth of British (learning French), German and Greek (both learning English) language learners aged between 14-15 years old using cross-

sectional data (N= 197). The British learners had approximately 210 hours of instruction while the German and Greek learners had more than double the amount of instruction amounting to 400 and 660 hours respectively. Learners' results on the receptive knowledge checklist test showed that after four years of instruction, the Greek learners came top learning 1,680 lemmas, German learners came second learning 1,200 lemmas and lastly were the British learners who learned only 660 lemmas. Similar to Webb and Chang's (2012) findings, the amount of instruction appears to be a key factor in vocabulary growth.

Laufer (Laufer, 1998) used cross-sectional data to compare the receptive and productive vocabulary growth of Hebrew EFL learners. She compared the test scores of grade 10 (16 years old; N = 26) and grade 11 (17 years old; N= 22) students at the end of the school year. The students received five hours of English instruction per week. She used the VLT (Nation, 1983, 1990) and a productive version of the VLT. The receptive vocabulary growth was 1600 word-families while the productive vocabulary growth was 850 word-families. This finding suggests that receptive vocabulary growth is likely to be larger than productive vocabulary growth.

In summary, this section reviewed the research on vocabulary breadth growth and highlighted the fact that there are only a handful of studies that tracked vocabulary growth using longitudinal data. The vocabulary growth of foreign language learners depends on a number of factors, a key one of which is the amount of foreign language instruction students receive where more input seems to lead to more vocabulary growth (Milton & Meara, 1998; Webb & Chang, 2012). Another key factor is the individual differences (e.g., motivation and out-of-class exposure) between learners which affect vocabulary growth as students in all previous studies differed in their vocabulary gains. For example, some

learners in Webb and Chang's study (2012) learned as few as 18 word-families while others were able to learn as high as 430 word-families. However, no research to date appears to have investigated this gap. The following section (section 2.3) discusses this and provides an overview of the role of individual difference in vocabulary learning.

2.3 Individual differences in vocabulary development

It has been shown in the previous section (section 2.2) that learners vary in their vocabulary growth. One of the main sources of variation in vocabulary and language learning is the individual differences between learners (Dörnyei, 2015; Dörnyei & Skehan, 2003; S. Li et al., 2022; Skehan, 1989). Individual differences are:

> "traits, dispositions, and characteristics, be they biological, social, psychological, or a combination of these, that make learners unique individuals, cause variation among learners, and are hypothesized to have a direct and/or indirect impact on learning outcomes" (S. Li et al., 2022, p. 4).

Individual differences in SLA serves as an umbrella term that includes several factors that have been found to influence language learning. These factors are commonly divided into three main categories: cognitive, conative, and affective (Cronbach, 2002). Cognitive factors refer to factors that influence language processing, storing and retrieval. The main factors under this category include language aptitude (Wen et al., 2019), working memory (Baddeley, 2003) and learning strategies (Oxford, 2017). Conative factors affect learners' goal-setting abilities and their abilities in persisting to achieve this goal. The major factor in this category is motivation, which has been researched extensively in SLA (Ushioda, 2020). Finally, affective factors influence learners' feelings and emotions which include

attitude (Mantle-Bromley, 1995), anxiety (Horwitz, 2001), enjoyment (Botes et al., 2022) and self-efficacy (C. Wang & Sun, 2020).

Ellis (2008) points out that the main objective of individual differences research in the past was to predict which learners will succeed in L2 learning. This was done to guide the selection of which learners are more fit to receive foreign language instruction. There has been a shift in the research objectives over the years and now researchers are mainly interested in explaining why some learners are more successful in L2 learning than others. This is pursued by analyzing the characteristics of the more successful learners with the practical aim of using these findings to guide learners on how to maximize their learning (for example, through teaching effective language learning strategies, Oxford, 2017).

Some individual differences (mainly learning strategies) have received substantially more attention in vocabulary research than other factors. For example, there is a full-length book (Takač, 2008) and book sections (Nation, 2022; Webb & Nation, 2017) on vocabulary and learning strategies, while little research exists on, for instance the role of self-efficacy in vocabulary development. It is only recently that there has been an increase in research aiming to provide an overview of the role of individual differences in vocabulary research (Dóczi & Kormos, 2015; Kim & Webb, 2022). Kim and Webb (2022) briefly reviewed the relationship between vocabulary knowledge and individual differences: working memory, aptitude, perceptual style, learning strategies, motivation, anxiety, previous L2 vocabulary knowledge, and age. They found more agreement on the relationship between vocabulary knowledge and some individual differences (e.g., vocabulary learning strategies, prior L2 vocabulary knowledge), while the effects of some factors (e.g., age) show more conflicting findings. Dóczi and Kormos (2015) focused on working memory, motivation and self-

regulation. They concluded that these factors have significant effects on vocabulary growth based on the reviewed studies. A key factor missing from both reviews is out-of-class exposure, which has emerged in vocabulary research in the past years, possibly due to the widespread access of learners to the internet and the emergence of smartphones and social media (Reynolds, 2023; Sundqvist, 2009; Sundqvist & Sylvén, 2016). This research area is gaining momentum but is still in need of further research (Kim & Webb, 2022; Schmitt, 2019). The three individual differences relevant to the focus of the present thesis are: out-of-class exposure, strategic vocabulary learning and motivation which the literature suggests to be key factors in vocabulary development (Dóczi & Kormos, 2015; Peters, 2018; Sundqvist & Sylvén, 2016). Each one of these three factors is discussed in detail in the following sections.

2.3.1 Out-of-class exposure

The classroom is not the only place where vocabulary development can occur. Some students come to class already knowing some English vocabulary (De Wilde & Eyckmans, 2017; Lefever, 2010). For example, De Wilde and Eyckmans (2017) found that a set of 11-year-old children in Belgium (N=30) could perform tasks at the A2 level (according to the Common European Framework of Reference for Languages) before receiving any formal instruction.

Out-of-class vocabulary learning can be categorized into extramural learning, extra-curricular learning and self-directed learning (Nation, 2022). Extramural learning (Sundqvist, 2009; Sundqvist & Sylvén, 2016) refers to learning from entertainment such as learning from watching television, playing video games, listening to songs and social

media interactions, none of which is under the control of the teacher. Extra-curricular learning (Benson, 2011) involves learning that is directed by the course or the teacher to supplement in-class learning. This can take the form of giving students a list of the target vocabulary in a course for them to learn intentionally at home. Self-directed learning (García Botero et al., 2019; Z. Li & Bonk, 2023; Nation & Yamamoto, 2012) is characterized by the learner taking full control of their own learning without the help of a language teacher during independent language learning or in conjunction with formal instruction. One example of self-directed language learning is learning a language from mobile-assisted language learning (MALL) apps such as Duolingo (Z. Li & Bonk, 2023). The type of out-of-class exposure investigated in this thesis falls under the category of extramural learning (hereon, "out of class exposure" will refer solely to extramural learning).

Several studies have found that out-of-class exposure has a positive effect on vocabulary development (Arndt & Woore, 2018; Feng & Webb, 2020; González Fernández & Schmitt, 2015; Peters, 2018, 2019; Peters & Webb, 2018). There is even some evidence from Peters (2018) that out-of-class exposure might have more effect on vocabulary learning than classroom instruction. She examined the relationship between gender, length of instruction (3 years vs. 6 years) and out-of-class exposure and receptive vocabulary knowledge. The results of the ANCOVA analysis showed that out-of-class exposure explained more variance (13%) than length of instruction (7%), while gender had no effect on test scores. The type of exposure (e.g., reading novels, TV viewing, listening to music and playing video games) might be an important factor in determining the quantity and quality of learning, however, most of the experimental research in this area has examined one type of

exposure at a time (Schmitt, 2019). The following sources of out-of-class exposure were selected for review since they are common sources of language input for many EFL learners around the world, as well as being relevant to the focus of the present thesis. The types of out-of-class exposure discussed are: extensive reading, extensive viewing of TV, listening to songs, playing video games and social media. Each is discussed to investigate whether these sources can lead to significant vocabulary learning and to discuss the factors that affect vocabulary gains.

2.3.1.1 Extensive reading

Extensive reading is the type of reading that students do primarily for pleasure and in large amounts. Bamford and Day (2004, p. 1) define extensive reading as:

> "an approach to language teaching in which learners read a lot of easy material in the new language. They choose their own reading material and read it independently of the teacher. They read for general, overall meaning and they read for information and enjoyment".

This is usually contrasted with intensive reading, which is the traditional reading conducted with the aim of learning language features such as grammar or vocabulary (Nation & Macalister, 2020). Extensive reading is perhaps the most researched type of out-of-class exposure with full-length books (Bamford & Day, 2004; Nation & Waring, 2020) and several journal articles (Al-Homoud & Schmitt, 2009; Nakanishi, 2015; Pigada & Schmitt, 2006; Stoeckel et al., 2012; Taguchi et al., 2004) published on this topic. What these studies tend to show is that extensive reading can lead to significant vocabulary learning (Day & Robb, 2015; Pigada & Schmitt, 2006; Suk, 2017). Extensive reading provides learners with

a large amount of comprehensible input, which is a necessary condition for SLA (Krashen, 1989; Rodrigo et al., 2004).

A meta-analysis including 34 studies and 3,942 learners found that extensive reading contributes to language development with a medium effect size of d = 0.46 (Nakanishi, 2015). Suk (2017) examined the effectiveness of extensive reading on vocabulary development over a 15-week school semester. 191 Korean EFL learners from four intact classes were assigned to two experimental groups and two control groups. Both the experimental and control groups had 100 minutes of in-class reading per week. The control groups received 100 minutes of intensive reading while the experimental groups received 70 minutes of intensive reading plus 30 minutes of extensive reading. In addition to in-class reading, students were asked to do out-of-class work. Students in the intensive reading classes were asked to do two to three hours of intensive reading and vocabulary exercises while the students in the extensive reading classes were asked to do two to three hours of additional extensive reading. The study used a self-made vocabulary test where the words were sampled from an extensive reading corpus. Results showed that the extensive reading classes made significantly more gains (13.07) than the intensive reading classes (3.41). One limitation of the study is the use of target words from an extensive reading corpus which might have favored the extensive reading group. Nevertheless, there is a large body of research that supports Suk's finding that extensive reading can indeed lead to vocabulary gains (P. Nation, 2022; Nation & Waring, 2020; Schmitt & Schmitt, 2020).

The gains from extensive reading (and incidental vocabulary learning in general) are usually small. Based on meta-analysis studies, the percentage of target words learned from incidental activities such as reading is 9-18% on immediate posttests and 6-17% on delayed

posttests. These rates gains are substantially smaller than the gains resulting from intentional vocabulary learning activities (e.g., flashcard learning) on immediate (18-77%) and delayed posttests (23-73%). Incidental vocabulary learning such as learning from reading involves less noticing and engagement with word forms which could explain the lower learning and retention rates (Laufer, 2003, 2005, 2010; Long, 1991; Schmidt, 1990). What this suggest is that students would need to read very large amounts of books to make substantial vocabulary gains (Cobb, 2007; Nation & Waring, 2020).

Extensive reading has become part of the language learning program of many language learning institutions (Stoeckel et al., 2012). Yet some learners and teachers might find the concept of extensive reading vague and might prefer more clear guidance. Day and Bamford (2004) suggest ten principles for effective implementation of extensive reading which provide guidance for both learners and teachers. The first five are relevant to out-of-class language learning (Day & Robb, 2015, p. 5):

- The reading material is easy (students are unlikely to enjoy a book if it is too difficult).
- A variety of reading material on a wide range of topics must be available (so students can find books they find interesting).
- Learners choose what they want to read (to enhance motivation).
- Learners read as much as possible (to make substantial gains).
- Reading speed is usually faster rather than slower (i.e., slow word-for-word reading might lead to poor comprehension).

Overall, most previous studies show that extensive reading can be an effective approach to vocabulary development. The gains are usually small, therefore, it needs to be done in large quantities.

2.3.1.2 Extensive viewing

In addition to traditional television, language learners today have unprecedented on-demand access to millions of online videos, TV shows and movies. YouTube for example has millions of videos and more than 500 hours of video are uploaded to YouTube every minute (Statista, 2022). These online videos can offer free, authentic, entertaining and informative content (Benson, 2015; Hung-chun Wang & Chen, 2020).

Several studies have investigated the effect of viewing audio-visual material (hereafter, viewing) on vocabulary learning and the common finding is that viewing can lead to lexical gains (Gesa & Miralpeix, 2023; Montero Perez et al., 2018a; Peters & Webb, 2018). For example, viewing was investigated by Peters and Webb (2018), where learners watched a one-hour documentary and their knowledge of 64 target words was assessed using meaning recall and meaning recognition tests. The findings showed that viewing resulted in significant incidental learning with word-related factors (frequency of occurrence and cognateness) and learner-related variables (prior vocabulary knowledge) affecting gains. Similar to extensive reading, extensive viewing can provide learners with ample amounts of comprehensible input (assisted by L1 and L2 subtitles, see next paragraph).

A number of potential factors that may influence lexical gains from viewing have been investigated. One key factor is subtitling (Baranowska, 2020; Frumuselu et al., 2015; Peters et al., 2016; A. Wang & Pellicer-Sánchez, 2022). The aim of these studies has been usually

1) investigating whether subtitles enhance vocabulary learning and 2) comparing L1 and L2 subtitles. The general findings emerging from these studies are that subtitles usually improve vocabulary learning and that L2 subtitles tend to lead to more vocabulary learning than L1 subtitles. For example, one of the early studies to show that viewing with subtitles leads to more vocabulary learning than viewing without subtitles is Koolstra and Beentjes (1999). They divided 246 bilingual 4th and 6th grade students into three experimental conditions: subtitles, no-subtitles and no-viewing (control). After watching a 15-minute documentary about grizzly bears, the subtitles group outperformed the no-subtitles group (on a written meaning recognition test and spoken form recognition test). Additionally, both viewing groups outperformed the control group. The majority of later studies have confirmed the advantage of viewing with subtitles compared to no-subtitles (Pujadas & Muñoz, 2019; Winke et al., 2010). The advantage seems to be due to subtitles helping language learners segment the speech stream (L2 subtitles only), guide their attention to unknown words and establish the form-meaning link (A. Wang & Pellicer-Sánchez, 2022; Winke et al., 2010). Meanwhile, there is less consensus on which subtitle type (L1 or L2) leads to more vocabulary learning. Most studies (Baranowska, 2021; Peters, 2019; A. Wang & Pellicer-Sánchez, 2022) and reviews (Reynolds et al., 2022; Wei & Fan, 2022) found that L2 subtitles lead to more vocabulary learning. For example, Frumuselu et al. (2015) asked university students with mainly (90%) Spanish/Catalan L1 background (other L1s included Dutch, German, Russian, Romanian and Moldavian) to watch the TV series 'Friends' over seven weeks. The 40 EFL participants were assigned randomly to either watch the show in L1 subtitles (Spanish) or in L2 subtitles (English). Results of meaning

recognition and meaning recall tests (15 each), showed the L2 subtitles group (English) significantly outperformed (posttest mean = 14.68) the L1 subtitles group (mean = 10.95).

On the other hand, some studies found no significant differences between the two types of subtitling (Bisson et al., 2014; Lwo & Chia-Tzu Lin, 2012; Muñoz et al., 2021). Muñoz et al. (2021) for instance examined the effect of subtitling from watching 24 episodes of a TV series distributed over an academic year. Results of mixed effects models showed no significant effect for subtitling on form and meaning recall tests. Pujadas & Muñoz (2019) suggest that the different results could be due to differences in research methodology (e.g., test modality, length of exposure) and learners' characteristics (e.g., proficiency and L1). In terms of proficiency, L1 subtitles might be more appropriate for beginning learners than L2 subtitles (Danan, 2004). This is based on the finding that understanding TV and movies (95% coverage) requires in part familiarity with the most frequent 3000 word-families (Webb & Rodgers, 2009a, 2009b). Empirical evidence gives support to this position. In an eye-tracking study, beginners with slow reading rates spent surprisingly very little time on each fixation[7] when L2 subtitles were used (Muñoz, 2017). Muñoz (2017) suggested that learners, due to their low proficiency, did not attempt to understand the audiovisual material.

Despite the mixed findings from the individual studies, results of a meta-analysis (Reynolds et al., 2022) and a review (Wei & Fan, 2022) on the topic suggest overall an advantage for L2 subtitles. One commonly provided explanation for the advantage of L2

[7] "The interval between the eye's movements, when the eyes 'stop', are called fixations." (Conklin et al., 2018, p. 30)

subtitles over L1 is that L2 subtitles can help language learners segment the speech stream, facilitating form-meaning mapping (Peters, 2019; Winke et al., 2010) which is a missing feature when L1 subtitles are used (Wei & Fan, 2022).

L1 and L2 subtitles are not the only types of subtitling. In some countries such as China, bilingual subtitles (i.e., where both L1 and L2 subtitles appear on the screen simultaneously) are widespread (M. Li & Hennebry-Leung, 2022; A. Wang & Pellicer-Sánchez, 2022). An eye-tracking study compared the eye movements and learning gains of 112 Chinese EFL learners in three conditions: L1 subtitles, L2 subtitles and bilingual subtitles (A. Wang & Pellicer-Sánchez, 2022). Form recognition, meaning recall and meaning recognition tests of novel target words were used. Results showed an advantage for bilingual subtitles over L2 subtitles in meaning recognition and over L1 subtitles in meaning recall. These advantages might be due to bilingual subtitles providing L1 meaning (facilitating access to meaning) and L2 form (facilitating attention to L2 form) simultaneously on the screen which might support establishing the form-meaning link. On the other hand, L2 subtitles were more effective in form recognition, possibly because the lack of another form of subtitling (i.e., L1) makes more attention resources available for learning L2 form. A similar advantage was found in another within-subject design study where students watched videos with L1 subtitles, L2 subtitles and bilingual subtitles (M. Li & Hennebry-Leung, 2022). After seven weeks of treatment, results of immediate and delayed tests (meaning recall and meaning recognition) showed an advantage for bilingual subtitles over L1 and L2 subtitles. Although current research on bilingual subtitles shows positive effects, it is still in the early stages. Both studies were conducted with intermediate to advanced learners therefore we are unsure if the same advantage applies to low-

proficiency learners. More research is needed to know if bilingual subtitles indeed bring the best of both worlds (of L1 and L2 subtitles) or merely introduce distraction to learners' limited cognitive resources, especially beginners (Wei & Fan, 2022).

In sum, viewing audio-visual input can lead to significant incidental vocabulary learning. There is a wide agreement that viewing with subtitles leads to more vocabulary learning than viewing without subtitles. Although there is less agreement on which subtitle type (L1 or L2) leads to more learning, results overall suggest an advantage for viewing with L2 subtitles, perhaps because it helps learners segment the speech stream (which is lacking in L1 subtitles), facilitating attention and learning of unknown words. Bilingual subtitles seem to be more effective than monolingual subtitles (L1 or L2 only) in learning meaning, yet further research is needed given the limited number of studies in this area.

2.3.1.3 Gaming

One key reason for the interest in the area of gaming and vocabulary learning is possibly due to the intrinsically motivating nature of playing video games (Nation, 2022; Zou et al., 2021). Boredom is one notable issue in foreign language classrooms (Kruk et al., 2021; Pawlak et al., 2020) and games offer a way of combating this by blending enjoyment with learning. It is important to establish first whether vocabulary learning can occur from playing games. Several studies have shown a positive correlation between the amount of video game playing and vocabulary knowledge (Brevik, 2019; H.-J. H. Chen & Hsu, 2019; De Wilde et al., 2019; De Wilde & Eyckmans, 2017; Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012). Sylvén and Sundqvist (2012) examined how the amount of time spent playing massively multiplayer online role-playing games

(MMORPGs) correlates with vocabulary knowledge. Questionnaires and diaries were used to measure the weekly amount of gaming and English language exposure (e.g., reading, viewing and listening to music) of young Swedish language learners (aged 11-12). Self-made tests of receptive (most frequent 1000 and 2000 levels) and productive vocabulary (2000 level) were used. Based on the amount of playing video games every week, students were divided into frequent gamers (five hours or more), moderate gamers (less than five hours) and non-gamers (none). Results of total vocabulary test scores showed that frequent gamers outperformed (vocabulary test mean = 25.4) moderate gamers (mean = 18.5) who in turn, outperformed non-gamers (mean = 16.6). One limitation of this study and previous research on the relationship between playing video games and vocabulary learning is that most research has been correlational which makes it difficult to establish causality (Field et al., 2012).

In response to this, some studies have used experimental approaches to investigate the effect of gaming on vocabulary learning (Aghlara & Tamjid, 2011; Cobb & Horst, 2011; Mohsen, 2016). For example, Mohsen (2016) randomly assigned 43 Arab adult students to either an experimental or control group. The experimental group engaged in a computer simulation game where they played the role of doctors performing knee surgery. The game involved following written instructions of the tasks to be completed (e.g., “Grab the sponge from the tool bar below so we can swab the leg with Betadine”). The control group only watched a video of the same surgery being performed. Following a pre and posttest design, results of meaning recognition tests (image association with words) showed that the experimental group (mean = 11.61) significantly outperformed the control group (mean = 7.90) on the posttest. Another study compared vocabulary learning (e.g., animal names)

from a video game to learning vocabulary using traditional methods (Aghlara & Tamjid, 2011). After a month and a half of instruction (90 minutes a week), the experimental group (mean = 7.8) significantly outperformed the control group (mean = 6.6) on a 10-item vocabulary test. Although no delayed posttests were used in both studies, the results of both show that playing video games can result in significant vocabulary learning.

When it comes to language learning through gaming, not all games are equally effective for vocabulary learning. Some games by default present more favorable conditions for learning than others (Reinhardt & Thorne, 2016; Sundqvist, 2019). Sundqvist (2013) proposed the Scale of Social Interaction model which hypothesizes that games where there is more interaction between the players will be more effective for language learning. Based on this model, MMORPGs are more effective for language learning than multiplayer games (since they often involve more interaction between the players), which in turn are more effective than single player games (Sundqvist, 2013). The model was tested in a study in which the relationship between the type of game learners played more frequently (MMORPGs, multiplayer games and single player) and their vocabulary knowledge was examined (Sundqvist, 2019). Results showed that learners playing multi-player games and MMORPGs had significantly larger vocabulary knowledge than learners playing single-player games. However, there were no significant differences between the vocabulary knowledge of learners playing multiplayer games and MMORPGs. As a result, the model was revised to take into account the non-difference between multiplayer games and MMORPGs (i.e., single-player < multiplayer games = MMORPGs). Although useful, the model is too specific and focuses only on the input coming from the players and neglects

the input coming from the game itself (e.g., the narratives and the conversations in the game).

Games can also be developed specifically for learning and training purposes (Johnson, 2007; Johnson et al., 2005). A common distinction is made between commercial off-the-shelf (COTS) games and serious games (H.-J. H. Chen & Hsu, 2019). Serious games are games that are designed primarily for learning (H.-J. H. Chen & Hsu, 2019; Johnson, 2007; Johnson et al., 2005). COTS games on the other hand are games designed mainly for entertainment and not learning. Although COTS games can result in vocabulary learning (Sundqvist, 2019), they might not be ideal for language learning due to linguistic and content factors (H.-J. H. Chen & Hsu, 2019). In terms of language, the primary audience of many COTS games (like many authentic sources of input) is native speakers which means that how vocabulary is treated might not be optimal for learning (e.g., including too many low frequency words, lack of repetition). Secondly, the content of some video games might not be appropriate in educational settings due to for example excessive violence. These factors led to the development of serious games which aim chiefly to educate but not at the expense of solid game design principles such as engaging game experience and immersing storylines (H.-J. H. Chen & Hsu, 2019; Kiili, 2005). Chen and Hsu (2019) examined vocabulary learning from a serious game that follows these guidelines. The game, Playing History, places the players in historical settings (e.g., one of its episodes is entitled The Slave Trade) and requires them to collect objects and complete missions. The game is suggested to be engaging, has rich language input and appealing storylines. 60 target words were selected: words that occurred only once were labeled low frequency, words with two to five occurrences were labeled intermediate and words occurring more

than six times were labeled high frequency. The same 60 words were used in a pre and posttest design with 66 university students in Taiwan (age mean = 19 years old). Results were organized by word frequency and showed that the largest gains occurred in the high frequency words (mean increase from the pretest = 28.36), followed by the intermediate frequency words (mean = 21.41) and finally the low frequency words (mean = 17.79). T-tests showed that all of these gains from the pretest to the posttest were significant ($p < .05$). The findings suggest that vocabulary learning can occur from playing serious games and that the amount of learning seems to increase as word frequency increases.

Overall, the findings from previous studies show that vocabulary learning can occur from both commercial off-the-shelf and serious games. More frequent gamers tend to have larger vocabulary size than less frequent gamers. It seems that games where there is more interaction between the players offer more opportunities for language learning than games where there is less interaction such as single player games. Finally, like other sources of incidental vocabulary learning (e.g., reading a book), words that occur more frequently are more likely to be learned.

2.3.1.4 Songs

It is perhaps more common to read a book or watch a movie once than multiple times, but this is not the case when listening to songs where repeated listening is the default (Abbott, 2002; Conrad et al., 2019). Repetition, as suggested throughout the thesis, is a key factor in vocabulary learning (Webb & Nation, 2017).

Songs are more similar to spoken language than written language and comprise mostly high frequency words (Romanko, 2017; Tegge, 2017). This makes songs particularly

useful for the learning of these words (Nation, 2022). Tegge (2017) examined two corpora, one consisting of 408 pop songs from US billboard charts and the other consisting of 635 songs selected by teachers for language learning purposes. The most frequent 3000 word-families provided 95.1% coverage of chart songs and knowledge of 6000 word families was necessary to reach 98.2% coverage. For the teacher-selected songs, knowledge of the most frequent 2000 word-families provided 95.5% coverage, while knowledge of the most frequent 4000 word-families provided coverage of 98.2%. These findings suggest that assistance is likely to be needed for understanding when listening to songs for beginners who have not mastered high frequency vocabulary.

Medina (1993) conducted one of the few empirical studies that have examined incidental vocabulary learning from listening to songs. She compared a story conveyed through song and the same story presented in a spoken format. Medina also examined the effect of using illustrations. The combinations of these factors resulted in four experimental conditions: narration, song, narration and illustration and song and illustration. Results showed no significant differences between the four conditions. However, the mean scores of the song group were higher than the narration group. Medina suggested based on the descriptive statistics that listening to songs may lead to vocabulary learning. She suggests that songs might provide learners with extra-linguistic support (rhythm) that might aid in word retention.

A more recent study was conducted by Pavia et al. (2019), who examined word learning (spoken form recognition and form-meaning link recognition) from listening to two different songs. The study also examined the effects of repeated listening to the same song (one, three or five times) and the relationship between frequency of occurrence (3-18) to

the target words and learning gains. The participants were 300 low level EFL students in Taiwan (Bilingual VLT score on the 1st 1000 frequency band = 8.57, 2000 band score = 4.39) aged between 10 and 14. There were eight groups in total: three listened to song A (one group listened once, another listened three times and the final group listened five times), three listened to song B (similarly, one group listened once, another listened three times and the final group listened five times) and two control groups. The data was collected in five 60-minute sessions each separated by a week. The results highlighted three key findings. First. listening to songs contributed to vocabulary learning yet the gains were small (0.52 words for song A and 1.64 words for song B, which is common in incidental word learning) and limited to spoken form recognition (i.e., not deep to the level of form-meaning learning; the authors hypothesized that this might be due to songs not having as informative context as other types of input such as reading). Second, repeated listening had a positive effect on vocabulary gains (the group who listened to song B five times outperformed other groups). Similarly, frequency of occurrence positively affected vocabulary learning. The authors recommend listening to songs both in-class and out-of-class as they appear to result in initial word learning (i.e., form recognition).

The fact that we tend to listen to the same song multiple times makes listening to songs theoretically a desirable input for vocabulary learning. Overall, the findings from the discussed empirical studies show that incidental vocabulary learning from listening to songs is possible and that repeated listening seems to lead to more vocabulary gains. Like other sources of incidental vocabulary learning, there is a need for a large amount of input before substantial gains are observed.

2.3.1.5 Internet and social media

There are nearly 4.8 billion users of social media every day which is approximately 60% of the world population (Ali, 2023). The average person spends more than two hours a day on these social media platforms (Ali, 2023). Social media is defined differently by different researchers. Reinhardt (2019, p. 1) defines social media as "any application or technology through which users participate in, create, and share media resources and practices with other users by means of digital networking". Major social media platforms include Facebook, YouTube, WhatsApp, Instagram, TikTok and X. Social media offer large quantities of authentic language input (listening to podcasts and reading blogs) and opportunities for language output (writing posts and speaking through engaging in online activities such as conversations and vlogs) which can help in language learning (Barrot, 2022).

There is little research (especially experimental) on the relationship between social media use and vocabulary learning (Nation, 2022). Some studies on overall out-of-class exposure include items regarding the frequency of visiting websites written in English and examine how they relate to vocabulary knowledge (De Wilde et al., 2019; Peters, 2018). For example, De Wilde et al. (2019) found that 78% of young language learners in Flanders (N = 780, aged 10-12) use social media in English daily. Results of their analysis showed that social media use had the highest correlation (r = .39) with the Peabody Picture Vocabulary Test (a test in which children match the spoken form of a word with a drawing representing its meaning; Dunn & Dunn, 2007) compared to other sources of out-of-class inputs (e.g., games, songs, TV). The use of social media also appears to help in the development of other aspects of vocabulary knowledge such as collocation (González Fernández &

Schmitt, 2015). González Fernández and Schmitt (2015) found that the use of social media as self-reported by the participants (0–1, 1–2, or more than 2 hours a week) correlated significantly with collocation knowledge test scores (form recall, r = .33).

Arndt and Woore (2018) conducted one of the few experimental studies on incidental vocabulary learning from social media. They compared L2 vocabulary learning (i.e., form, meaning and grammatical function) from written blog posts and video blogs (both had the same script). In this online experiment, the video group (n = 38) watched three vlogs while the blog group (n = 42) read three blog posts. Both the videos and the blog posts included the same six nonwords each occurring 11–14 times. Each target word was tested on written form recall, meaning recall, grammatical function recall, grammatical function recognition and meaning recognition. Results of the posttests showed that both the video (total vocabulary gain = 20.77) and blog groups (total vocabulary gain = 19.76) learned the nonwords without significant differences in total gains. In terms of vocabulary knowledge aspects, the two groups differed only in form recall (i.e., spelling) in which the blog group scored significantly higher. This result is expected since the blog group saw the written form of the nonwords during reading while the video group did not. The study used written tests which provide little details about how the two media differ in spoken vocabulary learning. Another limitation is the lack of delayed posttests, which hinders assessment of long-term vocabulary retention. The findings overall suggest that incidental vocabulary learning can occur from social media content whether this is in text or video format.

This section on out-of-class language exposure has shown that incidental vocabulary learning can occur from extensive reading, extensive viewing, playing video games, listening to songs and visiting social media platforms. One caveat is that sizeable gains will

only be possible when there is a large amount of out-of-class exposure (Milton, 2008; Nation, 2022; Schmitt & Schmitt, 2020). This will likely require both motivation (to initiate and maintain–out-of-class exposure) and self-regulation skills (e.g., to find and evaluate different types of out-of-class inputs, see next sections; Lai et al., 2015; Richards, 2015; Sundqvist & Sylvén, 2016). However, it is clear that research on vocabulary growth would be missing the full picture if it does not take into account what learners do beyond the classroom walls.

2.3.2 Strategic language learning[8]

The research on language learning strategies (LLSs) has expanded considerably following Rubin's study (1975) on the good language learner, with researchers aiming to define, classify and measure LLSs (O'Malley et al., 1985; e.g. O'Malley & Chamot, 1990; Oxford, 1990; Rubin, 1981; Wenden, 1991).

Oxford's (1990) volume on LLSs is one of the main studies in this area in which she defined and categorized LLSs, and constructed an instrument for LLSs assessment. Oxford (1990) defined LLSs as "specific actions taken by the learner to make learning easier, faster, more enjoyable, more self-directed, and more transferrable to new situations" (p. 8). Her taxonomy (1990) was one of the most widely used taxonomies of LLSs where she classified LLSs into six groups: Memory strategies (e.g., using keywords to remember words), Cognitive strategies (e.g., reasoning and summarizing), Compensation (e.g., guessing from context), Metacognitive strategies (setting goals and objectives), Social

[8] The term strategic learning is used here to describe the general construct of strategic knowledge approached through language learning strategies or differently through the concept of self-regulation (Tseng et al., 2006).

strategies (e.g., asking for clarification) and Affective strategies (e.g., lowering anxiety). The Strategy Inventory for Language Learning (SILL) is the most widely used instrument in LLSs research. The SILL has 50 self-report items corresponding to the six strategies mentioned earlier and uses 5-point Likert-scale responses ranging from "never or almost never true of me" to "always or almost always true of me". The SILL was used by Green and Oxford (1995) to examine the relationship between language proficiency and LLSs. In this study, 374 EFL participants from the University of Puerto Rico were divided into three proficiency levels. Results showed that the more proficient learners used LLSs significantly more frequently and diversely than the less proficient ones. Similar findings have been reported in other studies (Rubin, 1975; Stern, 1983; Wharton, 2000). In addition, strategies that involved active use of language in naturalistic settings such as watching TV in English or seeking opportunities to speak in English were used more often by the more proficient learners. These findings suggest a significant relationship between LLSs and language proficiency.

In an attempt to help the less proficient learners develop their language skills through LLSs use, strategy instruction (or training) was investigated. The findings from the different studies however are not straightforward. A critical appraisal of the literature by Rees-Miller (1993) found little success in strategy instruction. He attributes this to cultural differences, different educational backgrounds, ages, beliefs of students and teachers about language learning and different cognitive styles. Others, however, cautiously suggest that strategy instruction seems to be effective when conducted over a longer period of time (Macaro, 2006). More positive results are found in a meta-analysis by Plonsky (2011) which included 61 studies and 6,791 learners. The study found a small to medium effect size (d =

.49) of strategy instruction on language proficiency, which according to the author compares well with the overall average effect size of d = 0.40 found in educational research (Hattie, 1987). Overall, findings are not conclusive that LLS instruction leads to more effective language learning.

With the turn of the century, a number of scholars voiced some concerns regarding the validity of research on LLSs (Dörnyei & Skehan, 2003; Dörnyei, 2005; Skehan, 1989). The strongest of these is Dörnyei (2005), who called for abandoning the concept of language learning strategies altogether and replacing it with the more general concept of self-regulation (discussed in the next section). The main issue Dörnyei observed with LLSs research is definitional fuzziness, which results in the difficulty of distinguishing between "engaging in an ordinary learning activity and a strategic learning activity" (2005, p. 164). Dörnyei also criticized how LLSs are categorized. For example, he criticized separating memory strategies from cognitive strategies in Oxford's taxonomy (1990), arguing that memory strategies should be classified as cognitive strategies based on what later research has shown (Purpura, 1999). Finally, the decline of learning strategies in the field of psychology and the rise of self-regulation is an additional argument put forward as an indication of how the earlier is unfit for scientific research and that the latter should be pursued.

Despite Dörnyei's criticism, the research on LLSs did not cease (Dörnyei, 2001; Dörnyei, 2015, p. 140; Griffiths, 2020; Rose et al., 2018). However, the continuation of research should not be regarded as an indication that all issues have been addressed, but should rather be an indication that there is room for both LLSs and self-regulation to advance our understanding of strategic learning (Griffiths, 2020; Pawlak, 2021). This is manifested for

example in Oxford's (2011) Strategic Self-Regulation Model of Language Learning which combines both concepts in one model.

2.3.2.1 Strategic vocabulary learning

Being a key component of language, vocabulary has received attention in the work of Oxford and other researchers on LLSs (Cohen, 1996; O'Malley & Chamot, 1990; Oxford, 1990, 2017). The importance of vocabulary, manifested for example in vocabulary learning strategies (VLSs) being the most frequently used strategies by language learners (Schmitt, 1997), has contributed to VLSs becoming a key research area. Studies on VLSs have generally followed the same directions as LLS. Some studies have attempted to develop taxonomies (Gu & Johnson, 1996; Nation, 2022; Schmitt, 1997; B. Zhang & Li, 2011). Others have focused on the relationship between language proficiency and VLSs (Ahmed, 1989; Fan, 2003; Gu & Johnson, 1996; Kojic-Sabo & Lightbown, 1999). For example, Gu and Johnson (1996) correlated the VLSs of 850 college students with a vocabulary size test and a general proficiency test. From a number of strategies that were developed from previous research, Self-Initiation (being proactive and learning relevant and interesting vocabulary) and Selective Attention (knowing which words to focus on) were found to be positive predictors of the general proficiency test. Both of these, along with Activation strategies (seeking opportunities to practice newly learned words), showed a small but significant positive correlation with the vocabulary size test ($r = 0.35$, 0.24 and 0.31 respectively). On the other hand, Visual Repetition strategy (writing words repeatedly to memorize them) was the most negatively associated with both tests ($r = -0.2$). In general, the study found that more proficient learners employed significantly more diverse strategies, which other studies support (Ahmed, 1989; Fan, 2003; Sanaoui, 1995).

Another major line of research is developing methods and instruments for the investigation of VLSs (see Takač, 2008 for an overview). There are two main VLSs questionnaires commonly used in the literature: Gu and Johnson (1996) and Schmitt (1997), both of which are based on Oxford's (1990) SILL. According to Tseng et al. (2006), the items in SILL focus on specific strategic behavior instead of more general strategic traits. As a result, the SILL scales are not cumulative and calculating mean scale scores is unjustifiable psychometrically (Tseng et al., 2006). Due to the issues with LLS research and its instruments, Tseng et al. (2006) proposed replacing VLSs with the concept of self-regulation borrowed from educational psychology. Self-regulation is defined as "the ways that learners systematically activate and sustain their cognitions, motivations, behaviors, and affects, toward the attainment of their goals" (Schunk & Green, 2018, p. 1). Tseng et al. (2006) created the Self-Regulating Capacity in Vocabulary Learning scale (SRCvoc; used in both thesis studies Chapter 4 and 5). SRCvoc aims to measure learners' self-regulating capacity of strategic learning, which is the driving force of LLSs use according to the authors. As suggested earlier, the instrument is based on Dörnyei (2001) in which he conceptualizes self-regulation as consisting of five components:

Commitment control	(Maintaining or increasing commitment to achieving goals)
Metacognitive control	(Controlling focus and reducing procrastination)
Satiation control	(Eliminating boredom and making learning more enjoyable)
Emotion control	(Managing emotions by lowering negative feelings and promoting positive ones)
Environmental control	(Harnessing the environment to promote learning)

Each component is measured by four items in the SRCvoc, which add up to 20 items in total. A sample of items include: “Once the novelty of learning vocabulary is gone, I easily become impatient with it” (satiation control), and “When learning vocabulary, I have special techniques to achieve my learning goals” (commitment control). The authors suggest that the instrument has a high reliability based on their study (mean Cronbach Alpha $\alpha = 0.78$ with no individual scale below 0.70). The study is discussed in more detail in the next section as it examined both self-regulation and motivation (section 2.3.3).

2.3.3 Motivation

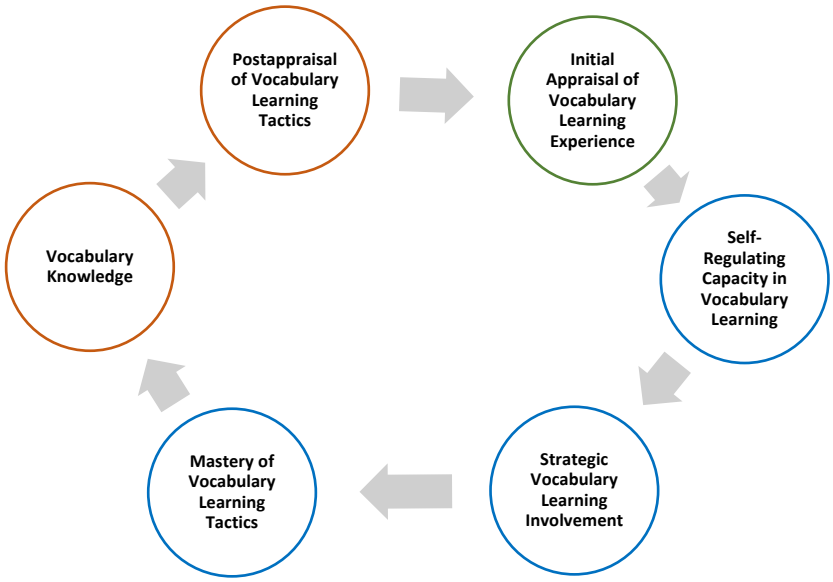
Language learners are expected to learn thousands of words to use language proficiently which is by no means a simple task. It takes them years of hard work and persistence to reach these numbers, and without motivation this goal would be unattainable. This is reflected in a number of studies that have found motivation to be a key factor in vocabulary development (Elley, 1989; Fontecha & Gallego, 2012; Gardner et al., 1985; Tremblay et

al., 1995; Tseng & Schmitt, 2008) and more general language skills (Jodai et al., 2014; Spolsky, 2000).

The research on motivation proceeded through three main stages (Boo et al., 2015; Dörnyei, 2015). The first stage was the social psychological period which emerged in the 1960s. It is commonly known for the integrative and instrumental types of motivation (Gardner & Lambert, 1972). This was followed by the cognitive-situated period in the 1990s which was marked by a move towards capitalizing on the advancements made in cognitive psychology by borrowing concepts such as Self-determination (Deci & Ryan, 1985) and Attributions (Weiner, 1992). Recognizing that motivation is a dynamic phenomenon led to the move to the process-oriented period, with the Process Model of L2 Motivation (Dörnyei & Otto, 1998) representing one of its seminal products. The Process Model of L2 Motivation sees motivation as composed of three stages: pre-actional (where motivation is generated), actional (where motivation is sustained and protected) and post-actional (where motivation is evaluated). Following a process-oriented perspective, Tseng and Schmitt (2008) used structural equation modeling (SEM) to examine the relationship between motivation, strategic learning and vocabulary development with six latent variables using questionnaires and vocabulary tests. The pre-actional stage is represented by the Initial Appraisal of Vocabulary Learning Experience (measuring vocabulary learning anxiety, vocabulary learning attitude and vocabulary learning self-efficacy). The actional stage is divided into Self-Regulating Capacity in Vocabulary Learning (measured using the SRCvoc discussed in the previous section), Strategic Vocabulary Learning Involvement (measuring the quantity of strategies used), and Mastery of Vocabulary Learning Tactics (measuring the quality of strategies used). Lastly, the post-actional stage

is represented by Vocabulary Knowledge (vocabulary breadth was measured using the VLT while vocabulary depth was measured using the overall score of collocation, polysemy and form recall tests of the words in the VLT) and Postappraisal of Vocabulary Learning Tactics (measuring the self-reflection of learning process phase after the learning task). The model was tested on 210 university students from China and Taiwan using a questionnaire and it showed in general a good fit with the experimental data. The best-fit model (Figure 2) depicted motivated vocabulary learning as being systematic and composed of cyclic and sequential stages (the learning process moves from one stage to the next). The authors suggest that the idea of cyclic learning aligns with the fact that learners typically need multiple encounters with a word to learn it. The model also suggests that motivation "is not just an "initial state" factor; it is an integral part of the whole system that drives the vocabulary learning cycle along" (Tseng & Schmitt, 2008, p. 383).

Figure 2. A structural equation model of motivated vocabulary learning (Tseng & Schmitt, 2008, p. 381)



A similar study was conducted by Zhang et al. (2017) but using different frameworks. To measure motivation, they framed their study within the theory of Self-determination (Deci & Ryan, 1985; Ryan & Deci, 2000), which conceptualizes motivation as orientations along a continuum from non-self-determined to self-determined. The theory makes key distinctions between intrinsic motivation (doing something because it's personally rewarding), extrinsic motivation (doing something to receive external rewards or avoid punishment) and amotivation (showing no motivation). Degrees of external motivation have been suggested along the continuum (Deci & Ryan, 1985) and Vallerand (1992) classified internal motivation into three types, which are shown in Figure 3.

Figure 3. Motivational orientations and the self-determination continuum

	Intrinsic motivation (types)	Knowledge Satisfaction from learning new information Accomplishment Pleasure from achieving goals Stimulation Excitement sensation from task engagement	
	Extrinsic motivation (degrees)	Integration	Integral sense of self
		Identification	Personal importance
		Introjection	Personal/imposed obligation
		External regulation	Gain reward or avoid punishment
	Amotivation	Showing no motivation	

Note. Based on the Center for Self-Determination Theory (2024)

To measure the self-determination of 107 Chinese learners, Zhang et al. used the Language Learning Orientations Scale (LLOS; Noels et al., 1999). This self-report questionnaire consists of 21 items, 9 of which target intrinsic motivation, 9 focus on extrinsic motivation and 3 on amotivation. The 7-point Likert responses range from 1 (not at all true of me) to 7 (very true of me). Instead of SRCvoc they used a VLS questionnaire adapted from Gu and Johnson (1996). The SEM in their study showed that VLSs function partially as a

mediator between motivation and vocabulary knowledge (i.e., VLSs translate motivation into actions that lead to vocabulary learning).

In the Saudi context, Alamer (2021a) constructed The Self-Determination Theory of Second Language (SDT-L2) based on Noles et al. (1999) to measure learners' L2 learning motivation. The SDT-L2 has 20 items with a 5-point Likert-type response format. The instrument is based on later research which suggests that two more global constructs of motivation subsume the four specific constructs discussed earlier (Ryan & Deci, 2017). The global construct of autonomous motivation includes intrinsic and identified orientations while controlled motivation includes introjected and external orientations. The instrument has two scales measuring the two main constructs and four subscales with 4 items in each one:

- Autonomous motivation
    - intrinsic
    - identified
- Controlled motivation
    - introjected
    - external

The SDT-L2 has been used in a study on 366 foundation-year Saudi students majoring in English from two universities (Alamer, 2021a). The results showed that autonomous motivation correlated positively and directly with vocabulary knowledge ($\beta = .23$, $p < 0.01$) while controlled motivation was negatively correlated ($\beta = -.25$, $p < 0.01$) with vocabulary knowledge. The instrument was further validated on 266 undergraduate Saudi students in

a study examining the relationship between motivation and L2 achievement measured by students' GPA scores in English (Alamer, 2021b). The SDT-L2 was found to be a reliable measure of L2 motivation based on model fit indices. The results also confirmed the earlier finding that higher levels of autonomous motivation correlate positively with L2 achievement, while controlled motivation correlates negatively with L2 achievement. Students acting on autonomous motivation are more likely to put more effort into their learning which tends to result in a stronger positive correlation with language learning outcomes (Alamer, 2021a; Dörnyei & Ushioda, 2011; Kormos & Csizér, 2014).

It is worth noting that autonomous and controlled motivations are not mutually exclusive, but it is possible for some learners to be driven by both at the same time (Liu & Oga-Baldwin, 2022; Vansteenkiste et al., 2009). For instance, Liu & Oga-Baldwin (2022) have identified three distinct motivational profiles among 523 Chinese EFL learners: High Quantity (high autonomous and controlled motivation), Moderate Quality (moderate autonomous motivation), Poor Quality (high controlled motivation). One study in a French university found that students driven by both high levels of autonomous and controlled motivation had similar academic achievement to students with high autonomous motivation (Gillet et al., 2017). This suggests that autonomous motivation might have the ability to mitigate the negative effects of controlled motivation (Liu & Oga-Baldwin, 2022).

The studies mentioned above are useful in improving our understanding of the relationship between vocabulary, strategic learning and motivation. Moving beyond the simple correlational studies and taking advantage of the potentials of SEM is a good step. However, the models need to be validated using different data (preferably from different

contexts) to assess their generalizability. Also, given the fluctuating nature of motivation, longitudinal data is likely to provide more accurate results.

The current section provided an overview of the role of individual differences in shaping the trajectory of vocabulary development. It examined three key sources of individual differences: out-of-class exposure, strategic learning and motivation. It highlighted their key role in explaining why some learners have larger and more developed vocabulary knowledge than others. The review also points to a key gap in the literature in that no previous study has examined how multiple individual differences jointly affect vocabulary growth using longitudinal data. This might be necessary since longitudinal studies provide a more accurate description of vocabulary development compared to cross-sectional studies (Schmitt, 2010, 2019), which only offer a snapshot of the relationship between individual differences and vocabulary knowledge (see study1 in Chapter 4 which measures vocabulary development longitudinally and while taking into account individual differences among learners).

3. An overview of English vocabulary learning, instruction and research in Saudi Arabia

The last chapter provided a general background about vocabulary learning and SLA more broadly. This chapter focuses on English vocabulary learning in Saudi Arabia where the two thesis studies were conducted. The chapter also bridges a gap in the literature given that there is little research that provide an overview of English vocabulary learning in the Saudi context.

English is the most common second language in the world, spoken by one in every four people (Beare, 2019). This is not surprising given that it is the primary language of key fields such as business, science and aviation (Crystal, 2017). Saudi Arabia is no different; English plays a pivotal role in the country due to educational, economic, social, cultural and historical reasons (Mahboob & Elyas, 2014; Moskovsky & Picard, 2018). In the context of education, the importance of English in Saudi Arabia is captured by the fact that it is the most learned foreign language, taught from first grade, and it is the medium of instruction in higher education. As discussed in Chapter 2, vocabulary is central to all English language use (Clenton & Booth, 2020) including listening (Y. Li & Zhang, 2019), speaking (Uchihara & Clenton, 2023), reading (D. D. Qian, 2002; S. Zhang & Zhang, 2022) and writing (Stæhr, 2008). Therefore, Saudi EFL learners need to have a sizeable English vocabulary to understand and produce English effectively.

The chapter starts by tracing early research on Saudi EFL vocabulary learning. Second, it provides an overview of the educational system and English instruction in Saudi Arabia. Third, it examines whether Saudi EFL students have enough vocabulary knowledge to understand everyday English and undertake more advanced tasks such as reading authentic and unsimplified English texts. Fourth, it examines the quality of vocabulary instruction in Saudi Arabia and teachers' awareness of effective vocabulary teaching. Finally, it critically examines previous research that has aimed to improve the vocabulary learning of Saudi EFL students.

3.1 A brief history of English vocabulary learning of Saudi students

Studies that involve vocabulary learning and Saudi EFL students date back to the late 1960s. Za'rour and Buckingham (1969) investigated the English language learning of Saudi learners and learners from other nationalities (e.g., Afghan, Iranian, Jordanian, Moroccan Turkish) learning English at the American University of Beirut in Lebanon. The authors aimed to investigate the role of individual differences and other factors (nationality, gender, L1 background, age, prior test scores and financial assistance) in English language learning. The participants were 239 male and 45 female students who joined a 15-week intensive English course. Their learning was measured by testing them at the beginning and at the end of the course. The 200-item multiple-choice test included four parts: vocabulary (50 items), structure (50 items), reading comprehension (50 items) and miscellaneous items (e.g., use of dictionary and correction of composition). The total scores revealed that the Saudi participants scored significantly the lowest both at the beginning and at the end of the course. Scores on the vocabulary part were not reported. The authors speculated that the low proficiency of Saudi students might be due to their

relatively lower contact with English speakers compared to other countries such as Jordan due to British colonization (Saudi Arabia on the other hand was not colonized by any European power; Al-Rasheed, 2010). Another early study, conducted by Zeiss (1983), examined the effect of music and relaxation on learning technical vocabulary. The participants were 14 Saudi students learning English as a second language in a US college. The results found no significant differences between the experimental group and the control group. Given that all students in the experimental group had perfect scores on the posttest, it is possible that there was a ceiling effect, which may have masked any potential benefits of the intervention by limiting the ability to detect further improvement.

Work focusing primarily on the vocabulary knowledge of Saudi EFL learners emerged in the early 1990s. The first of these[9] was conducted by Al-Hazemi (1993) who examined the vocabulary size of secondary students enrolling in a military academy (aged 19-23). He used a checklist test (Eurocentres Vocabulary Size Test; Meara & Jones, 1990) and found that Saudi EFL students had an average vocabulary size of 1000 word-families. There are no published reviews on vocabulary learning and teaching in Saudi Arabia, hence the current chapter aims to address this as well as providing background knowledge about the context of the thesis studies.

3.2 An overview of English education in Saudi Arabia

The educational system in Saudi Arabia consists of five stages: kindergarten, elementary (grade one to six), intermediate (grade seven to nine), secondary (grade ten to twelve) and

[9] Searching with the terms ‘Saudi’ and ‘vocabulary’ in Google Scholar revealed no studies that focus primarily on vocabulary and Saudi learners before Al-Hazemi (1993).

higher education (Elyas & Picard, 2018). Children typically start school at the age of 5-6 and graduate from secondary school at the age of 17-18. Public schools are segregated by gender in which male students (grade four and higher) are currently taught exclusively by male teachers while female students are taught exclusively by female teachers. Arabic is the official language of public education (up until secondary education) while English is the medium of instruction in higher education programs (e.g., Math, Medicine and Chemistry are all taught in English). The educational system is very centralized with standardized curricula and textbooks (Al-Hoorie, Al-Shahrani, et al., 2021). As a result, teachers have little freedom in teaching content beyond what is prescribed in the textbooks.

Formal English language teaching in Saudi Arabia dates back to 1928 according to Al-Seghayer (2011), but it was not until 1945 that English was taught in secondary schools (Al-Hoorie, Shlowiy, et al., 2021). The grade at which English is introduced to students has undergone significant changes in the last 20 years (Barnawi & Al-Hawsawi, 2017). Pre-2004, it was limited to intermediate and secondary levels. However, in 2004, it was expanded to include sixth-grade students. By 2011, it reached fourth-grade level. Presently, as of 2021, English instruction commences as early as first grade (age 5-6). Saudi students have less instruction time compared to neighboring countries such as Oman and Kuwait in which English is taught five times per week from elementary grade (Alsuhaibani et al., 2023). English is currently taught three times per week in elementary school, four times a week in intermediate school and mostly five times per week in secondary school (Ministry of Education, 2023). English classes span 45 minutes each, resulting in elementary students receiving 2 hours and 15 minutes of English instruction per week. Intermediate students get 3 hours, while secondary students receive 3 hours and 15 minutes weekly. By the time

Saudi students leave secondary school, they will have accumulated over 1000 hours of formal English instruction (weekly hours x 38 school weeks a year).

Students in Saudi Arabia are typically taught English by native Arabic speaking teachers (Moskovsky & Picard, 2018) who tend to rely heavily on Arabic in the classroom (Alshammari, 2011). Alshammari (2011) found that 60% of Saudi EFL teachers believed using Arabic was necessary to save time, while 69% used it to clarify difficult concepts. The issue of overreliance on Arabic is compounded by the fact that many teachers dominated classroom discourse, providing limited opportunities for student engagement (Alsaedi, 2012). Students' low engagement is not only due to teachers' dominance but could also be attributed to students' reluctance to engage in communicative activities (Aljumah, 2011). Farooq (2015) recognized the potential of communicative language teaching (CLT), which is an approach to language teaching that emphasizes meaningful and real language use as being central to language learning (Richards & Rodgers, 2014), for improving learners' communicative competence. However, he identified barriers to CLT implementation, such as overcrowded classes, lack of resources, and low learner proficiency levels.

The English education system in Saudi Arabia has undergone significant transformations in recent years, particularly in terms of expanding the introduction of English to younger grades and increasing instructional hours. However, the centralized nature of the curriculum, limited instructional time, and reliance on traditional teaching methods seemingly continue to hinder students' language proficiency. Teachers' overreliance on Arabic and limited use of communicative approaches appear to further restrict student

engagement and communicative competence (Study 1 in Chapter 4 provide insights into the vocabulary development of Saudi EFL students).

3.3 The vocabulary knowledge of Saudi EFL students

A survey of the literature on Saudi EFL students' vocabulary knowledge shows only four studies conducted with secondary students (Alhaj et al., 2019; Al-Hazemi, 1993; Alsaif, 2011; Alzahrani, 2020) and none primarily with intermediate students. On the other hand, there are several studies conducted with university students (Al-Homoud & Schmitt, 2009; Al-Khasawneh, 2019; Al-Masrai & Milton, 2012; Al-Nujaidi, 2003; Alqarni, 2019; Alsharif, 2022; Altalhab, 2019). Saudi researchers, like SLA researchers more generally, focus predominately on college and university students. Several researchers have voiced concerns regarding this bias in sampling which might skew our understanding of SLA and language teaching (Andringa & Godfroid, 2019, 2020; Cox, 2019; Ortega, 2019). One evaluation of this skewness is Plonsky (2016) who examined 600 studies from six SLA journals and found that approximately 67% of the participants were college or university students. Younger language learners in particular are underrepresented in SLA research despite the fact that they are larger in number than university students (Kormos & Sáfár, 2008). In Saudi Arabia for instance, there are more than six million young language learners (in k-12 education) while the number of students in higher education (university and college) is approximately 1.4 million (Fawaaz, 2023). Andringa and Godfroid (2020) aimed to gauge this bias by examining 17 meta-analyses from SLA research and found that young language learners account for only 25% of the research samples. What this suggests is that there is a need for more balanced sampling, including more research on younger language learners, both in the Saudi context and in SLA in general.

Table 3. Receptive vocabulary knowledge of Saudi EFL learners

Author	Test	Participants		Vocabulary scores mean	Notes
		N	Level		
AL-Hazemi (1993)	Checklist	137	Military academy	1000	Military cadets
Alsaif (2011)	Checklist	139	Secondary	890 (lemma)	5000 ceiling
Alhaj et al. (2019)	VLT	80	Secondary	1000	2000 ceiling
Alzahrani (2020)	VST	108	Secondary	2025	4000 ceiling
Al-Nujaidi (2003)	VLT	226	University 1st year	500-700	3000 ceiling
Al-Khasawneh (2019)	VLT	64	University 1st year	2025	English major
Alamer (2021)	VLT	366	University	2086	5000 ceiling
Alqarni (2019)	VLT	71	University Final year	Males: 2435 Females: 1990	English major
Altalhab (2019)	VST	120	University 1st year	3000	English major

Alsharif (2021)	VST	116 female	University	3871	5000 ceiling & English major
Masrai & Milton (2012)	Checklist	55 & 37	University 1st year & final year	1st year: 1680 Final year: 4198	English major

Notes. Vocabulary knowledge is reported in word-families except for Alsaif (2011) who used a lemma-based test

The Ministry of Education in Saudi Arabia expects school students to know 3000 word-families by the time they graduate from secondary school (Al-Masrai & Milton, 2012). The goal is reasonable since it provides learners with the minimal lexical coverage (95%) needed to understand spoken discourse and everyday conversations (van Zeeland & Schmitt, 2013). It is clear based on the Table 3 above that most studies find that Saudi secondary students (and even university students) are below this number. Alsaif (2011) examined whether secondary school students do achieve this number upon graduation using a checklist test and found they fall far below expectations. His results showed that students graduate from secondary school with a mean score of 890 lemmas (roughly 556 word-families; lemma/1.6 = word-families, Milton, 2009). Alhaj et al. (2019) examined the vocabulary knowledge of secondary students (aged 16-19) using an older version of the VLT (VLT version not reported). They found that Saudi secondary students know approximately 1000 word-families. The two studies, nearly a decade apart, show not much improvement in the vocabulary knowledge of Saudi EFL students over the years. More optimistic results come from Alzahrani (2020). He examined the vocabulary knowledge of

108 secondary students and found that they know around 2025 word-families which is approximately double the two previous studies. It is possible that the sample in his study included more motivated learners or more qualified teachers, both of which are factors that can lead to more vocabulary learning (Webb & Nation, 2017). Overall, the studies show that secondary students know between 500-2000 word-families which is below the 3000 word-families goal. This means that Saudi secondary students will likely face serious difficulties understanding basic and everyday English since they lack knowledge of many high frequency words.

This low vocabulary knowledge also extends to university students as shown in the previous table (Table 3). It shows that the vocabulary knowledge of Saudi university students ranges from 500 to 4198 word-families. However, most of the sampled studies show a vocabulary knowledge in the range of 2000-3000 word-families. One issue with previous research on the vocabulary knowledge of Saudi university students is recruiting English major students. These students are not the most representative of the actual vocabulary knowledge of university students since they often receive more language instruction than students in other majors due to their specialized studies (see section 2.2.4 on the effect of the amount of input on vocabulary learning). In the Saudi context for instance, most English language programs dedicate a whole year (the second year after the first preparatory year) to intensive English courses that target the four skills (listening courses, reading courses, speaking courses and writing courses) in addition to grammar courses (Al-Hoorie, Al-Shahrani, et al., 2021). Students in other majors do not normally receive this extra year of English instruction. Therefore, studies should ideally aim to

sample from different majors if possible or at least from a non-English major sample for more representative vocabulary size estimates (Mclean et al., 2014; Ngoc Yen, 2020).

Another issue in the studies that aim to measure the vocabulary size of university students is using a VLT instead of a vocabulary size test (5 of the 11 studies in Table 3). VLTs were not intended for measuring vocabulary size; they were designed to be diagnostic tests that help teachers assess which frequency band (e.g., first or third 1000) students need to focus on (see section 3.4.2; Nation, 2022; Webb & Nation, 2017).

In general, more reliable vocabulary size estimates of university students come from studies that use a vocabulary size test (e.g., the VST), test sufficient bands to avoid a ceiling effect (some studies include only a few levels from the VST, see Table 3) and sample from non-English major students (to better represent the majority of university students). None of the reviewed studies meet all of these requirements. For example, most of the studies in Table 3 have English major participants (Al-Masrai & Milton, 2012; Alsharif, 2022; Altalhab, 2019).

Nevertheless, these studies can provide some general estimations of Saudi EFL students' vocabulary knowledge. Al-Nujaidi (2003) conducted one of the earliest studies on the vocabulary knowledge of university students. He used a VLT test with 226 first year university students and found that they knew 500-700 word-families. The study has the limitation of using a VLT as a vocabulary size test and being relatively outdated (e.g., English was not taught at the elementary level at that time). Al-Masrai and Milton (2012) examined the vocabulary knowledge of 92 first and final year students majoring in English. Their findings showed that students know between 2000 to 3000 words when they join the

English program and graduate with an average of around 5000 words. A more recent study shows similar vocabulary sizes for first year students majoring in English (Altalhab, 2019). The vocabulary size of 120 first year English major students was measured using the VST and the results showed a vocabulary size of 3000 word-families. Alsharif (2022) examined the vocabulary size of English major female students (mean age = 23.31, SD = 4.45). The findings showed that these students have an average vocabulary size of 3871 word-families. The reviewed studies show that English major university students in Saudi Arabia seem to know on average 3000 word-families during academic studies and appear to graduate with a vocabulary size of 5000 word-families. Recalling that 6000 to 7000 word-families are needed for unassisted listening and 8000 to 9000 word-families for reading (Nation, 2006), these learners are unlikely to be independent language users. That is, their vocabulary size provides a minimal comprehension of spoken and written language (95% coverage) but not an optimal one (98% coverage) (i.e., they are likely to encounter several unknown words that might hinder comprehension; Laufer & Ravenhorst-Kalovski, 2010). The vocabulary size of non-English major university students is likely to be lower given that they receive less English instruction compared to English major students, yet little is known due to a lack of studies that look at a broad range of participants in a systematic way.

The chapter so far has shown that Saudi EFL students in general have limited vocabulary knowledge. A number of different factors have been proposed to explain this low vocabulary and language proficiency, such as low vocabulary input from textbooks (Alsaif & Milton, 2012), low autonomy of Saudi EFL students (Alrabai, 2017) and the lack of learning resources in many schools (e.g., language labs and English books; Almutairi,

2008). Factors more relevant to the focus of this thesis include students' limited exposure to English out-of-class (Al-Homoud & Schmitt, 2009; Moskovsky & Picard, 2018). One form of out-of-class exposure is extensive reading which, as noted by Al-Homoud & Schmitt (2009), is not very common in Saudi Arabia. Another relevant factor is the low motivation levels of Saudi students as reported in several studies (Alrabai, 2016; Moskovsky et al., 2013). Given the dominance of Arabic, many Saudi students do not see an immediate value in learning English (Alrabai, 2018). Finally, some studies such as Alqarni (2018) cite the limited use of vocabulary learning strategies by Saudi students as a factor for their small vocabulary. Alqarni used a self-made questionnaire based on Schmitt's (1997) taxonomy of vocabulary learning strategies and found that the overall mean of vocabulary strategy use was 1.63 (on a five-point scale ranging from never to always). Alongside these important factors is the role of instruction and teachers in students' vocabulary learning which is discussed in the next section.

3.4 Vocabulary instruction in Saudi Arabia

For the majority of EFL learners around the globe, the classroom is the primary source of language learning. Thus, vocabulary researchers over the years have produced practical findings that can be applied directly to the language classroom. These areas include for example word selection (identifying which words are most useful to learners), enhancing learning (e.g., using glossing in reading to draw learners' attention to target words) and improving retention (e.g., word recycling; Nation, 2011, 2022; Webb & Nation, 2017). The following discussion will present some of these recommendations before discussing vocabulary instruction in Saudi Arabia.

Since classroom time is limited in many places and the number of words learners need to know is in the thousands (Lightbown & Spada, 2020), teachers need to be careful when selecting which vocabulary to focus on. The usefulness of a word is usually determined by its frequency, with high frequency words being deemed more valuable since learners will encounter them more often than less frequent words (Nation, 2022). In addition to vocabulary selection, vocabulary researchers suggested a number of conditions that promote vocabulary learning (Webb & Nation, 2017). These conditions fall into two main categories (Webb & Nation, 2017): repetition (i.e., how many times a word is encountered) and quality of attention (i.e., depth of word processing). The likelihood of vocabulary learning increases the more a word is encountered (Pellicer-Sánchez, 2016; Uchihara et al., 2019; Webb, 2007b) and the more deeply it is processed (e.g., successful retrievals of word form or meaning, meeting and using a word in different contexts; Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2008; Nation & Webb, 2011). In terms of repetition, research suggests that it is rarely the case that word learning can occur from a single encounter (Webb, 2007). In a meta-analysis involving 1,918 participants from 26 studies, a medium effect ($r = .34$) was found for repetition on incidental receptive vocabulary learning (Uchihara et al., 2019). The question of how many repetitions are needed to learn a word has been investigated, with studies reporting learning on a recognition test from 10 encounters in a short text (Webb, 2007b) to well over 20 in a longer text (3 graded readers; Brown et al., 2008). However, it is unlikely that there will be a fixed number for all words and in all contexts given the complexity of this seemingly simple question (Laufer & Rozovski-Roitblat, 2015).

In terms of the quality of attention (Webb & Nation, 2017) or engagement (Schmitt, 2008), the more deeply a word is processed, the more likely it is to be learned (H. chao M. Hu & Nassaji, 2016; Hulstijn & Laufer, 2001; Nation & Webb, 2011). The idea of deeper processing and more engagement has undergone different conceptualizations, the two most influential of which are the Involvement Load Hypothesis (Hulstijn & Laufer, 2001) and the Technique Feature Analysis (TFA) (Nation & Webb, 2011). The ILH suggests that engagement with words has three components: need, search and evaluation. Need refers to the condition that a word is needed to complete a task (e.g., fill in the blanks). Search refers to whether information about a word is provided (e.g., definition of a word in a text is made available) or whether the learner should find this information (e.g., looking up the meaning of a word in a dictionary). Evaluation refers to whether a learner is required to assess if a word fits in a specific context or that such requirement is low. Vocabulary learning activities are more likely to be conducive to learning when there is high need, search and evaluation and less likely when these components are moderate or weak. The ILH's ability to predict effective vocabulary learning activities has been evaluated in a meta-analysis that examined 42 empirical studies involving 4,628 participants and 398 effect sizes (Yanagisawa & Webb, 2021). Results demonstrated that the ILH significantly predicted learning and explained 15% (immediate posttests) to 5.1% (delayed posttests) of the variance. Among the ILH components, evaluation contributed most to learning, followed by need, while search showed no significant impact. One limitation of the ILH is that it leaves the learner out of the equation (Schmitt, 2008). The TFA was developed to address this by being more inclusive than the ILH through the inclusion of factors such as learners' motivation. Similar to the ILH, the TFA framework evaluates the effectiveness of

vocabulary learning activities. It covers five main components, each containing between 3 to 5 criteria:

- Motivation (e.g., ‘Does the activity motivate learning?’)
- Noticing (e.g., ‘Does the activity focus attention on the target words?’)
- Retrieval (e.g., ‘Are there multiple retrievals of each word?’)
- Generation (e.g., ‘Does the activity involve generative use?’)
- Retention (e.g., ‘Does the activity involve imaging?’)

The criteria are scored dichotomously (1 = feature exists, 0 = feature is missing), with a higher score (max 18) indicating a more effective vocabulary learning activity. The predictive capacity of the ILH and the TFA were compared in a study (Hu & Nassaji, 2016). 96 adult EFL learners were divided into four groups and tasked with learning 14 unknown words using different vocabulary learning methods based on these frameworks. The findings indicated that the TFA demonstrated greater predictive capability for vocabulary learning gains compared to the ILH. One of the tasks that led to more learning involved a productive component which the TFA takes into account (i.e., generation) but not the ILH. These findings suggest that the TFA might be more effective in facilitating effective L2 vocabulary learning tasks than the ILH.

Although both repetition and quality of attention are key to vocabulary learning, the latter might be a more influential and stronger predictor of learning (Eckerth & Tavakoli, 2012; Laufer & Rozovski-Roitblat, 2015). For example, Laufer & Rozovski-Roitblat (2015) examined both repetition (1 to 21 encounters) and quality of attention in three tasks with varying levels of engagement (reading only, reading with a dictionary, reading and word-

focused exercises) over 11 weeks among 185 learners. Results of the posttests showed that reading combined with word-focused exercises (which has the most engagement) led to the best outcomes regardless of the number of encounters. The authors conclude that what learners do when they encounter words is more important than the number of times they encounter them.

Students' vocabulary growth will be in part influenced by the application (or lack of) of vocabulary research recommendations discussed above. Teachers play an important role in the classroom and their approach to vocabulary teaching (e.g., their verbal lexical explanation of word meaning) influences students' vocabulary development (Dang & Webb, 2020; J. H. Lee & Lee, 2022). Sonbul et al., (2022) examined the awareness of effective vocabulary instructional practices of EFL teachers in Saudi Arabia. The participants were school and college-level EFL teachers who completed a survey (n = 86) and a focus group interview (n = 15). The survey items were formed around the criteria in the TFA framework. For example, one of the survey questions is based on one of the items in the TFA noticing criteria: 'When I design a vocabulary activity, I make sure that it focuses attention on the target words'. The five-point response ranged from 'never' to 'usually'. Results showed that teachers in Saudi Arabia pay more attention to some criteria than others. They focus more on setting clear goals, motivating students and raising students' awareness of new word learning. In contrast, they pay less attention to spacing retrievals (i.e., distributing the learning time over multiple sessions, e.g., learning 10 minutes a day for a week instead of learning 70 minutes at once in a single day) or giving learners the freedom to select the words to learn. In terms of the differences between school and college teachers, the findings of the survey and the interviews showed higher

sensitivity by college teachers to the lack of freedom to choose target words in the Saudi curriculum. This might indicate a higher awareness of the fact that textbook-assigned words might not be readily suitable for learners and that students might need support with other vocabulary (e.g., high or mid-frequency). Meanwhile, college teachers preferred defining unknown words in English - something not supported by vocabulary research, which tends to show higher retention with L1 definitions (Masrai & Milton, 2015; Webb & Nation, 2017). Overall, the findings of Sonbul et al. (2022) suggest that school and college-level EFL teachers in Saudi Arabia seem to lack a solid background knowledge in how to teach and guide vocabulary learning effectively inside the classroom. The authors recommend that teachers in Saudi Arabia should receive specialized training courses to increase their awareness of best practices in vocabulary teaching.

3.5 Research on improving Saudi EFL students' vocabulary knowledge

Several studies have aimed to improve the vocabulary learning of Saudi EFL students. Although general vocabulary research might be applicable to the Saudi context, the ecological validity and relevance of vocabulary research conducted specifically with Saudi students is likely to be higher since it better reflects more the characteristics of this group (i.e., Saudi EFL students; Vu & Peters, 2021). One way of classifying these studies is by categorizing them into research focusing on intentional vocabulary learning and research focusing on incidental vocabulary learning. As discussed in section 2.1.5., intentional learning (e.g., flashcard learning) usually leads to more vocabulary gains than incidental learning (e.g., picking up words from reading). However, intentional learning has the limitation of not being effective for aspects beyond the form-meaning link (e.g., collocation and register) since teaching all other aspects for every word is likely to be time-consuming.

On the other hand, incidental vocabulary learning tends to lead to small gains but allows other aspects of vocabulary knowledge such as collocation knowledge to be picked up (Webb et al., 2013). The view held by most vocabulary researchers is that these approaches are complementary and should be included for balanced vocabulary learning (Durrant et al., 2022; Nation, 2007; Schmitt & Schmitt, 2020).

One of the studies on intentional vocabulary learning was conducted by Sonbul and Schmitt (2009) in which they compared reading only to reading plus explicit teaching of words (writing words on the board and explaining their meanings). The 40 Saudi university students in the study were tested on form recall, meaning recall and meaning recognition. The items were 20 low-frequency words mixed with 40 high-frequency words. Results of immediate and delayed tests showed significantly higher gains for the reading plus group on all tests. The authors suggest that explicit instruction can lead to deep learning as evidenced by students scoring higher on the form recall test (which is the most difficult aspect of form-meaning link knowledge; González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004). Especially for high-frequency words, the results of this study show that bringing students' attention to words through direct instruction can lead to more significant gains.

Another way to effectively boost the vocabulary knowledge of students is the use of flashcards. A meta-analysis comparing a number of intentional learning activities (flashcards, wordlists, fill-in-the-blanks and writing sentences) found that flashcard use was the most effective as measured by effect size (Webb et al., 2020). Flashcards have been employed by Sanosi (2018) to help low-level Saudi university students expand their vocabulary. The experimental group (n = 21) learned 90 words in-class through Quizlet (a

digital flashcard learning platform) for a month while the control group (n = 21) learned vocabulary following traditional approaches. Results of the posttests (multiple choice, gap-filling and matching) showed that the group who learned from flashcards significantly outperformed the control group. The results confirm the well-established finding that flashcard learning can lead to significant vocabulary gains (McLean et al., 2013; Nakata, 2020). Relevant to flashcard learning is the importance of spacing the learning of vocabulary for more retention. In the Saudi context, Alfotais (2019) compared spaced and massed learning (i.e., cramming) across the four form-meaning link aspects (meaning recognition, form recognition, meaning recall and form recall). Following a within-subject design, first year university students (n = 62) learned 30 words once in a spaced condition and another in a massed condition in authentic EFL classroom settings. The results of the immediate and delayed posttests showed significantly higher gains and retention in the spaced learning condition. The results revealed that the differences were higher in recall tests (mean difference > 3) than in meaning recognition tests (mean difference < 1.5). The findings highlight the importance of spacing for more vocabulary learning and retention.

Based on the finding that individuals tend to retain information more effectively when presented with both visuals and texts compared to texts only (e.g., Mayer, 2001), Alwadei and Mohsen (2018a) investigated how vocabulary learning in the classroom might be improved through the use of infographics (i.e., graphics that aim to simplify complex information). 41 Saudi EFL learners were divided into an experimental group (learning words from infographics) and a control group (learning through traditional instruction). After 10 weeks of learning, the results of the immediate and delayed tests (4 weeks) of meaning recall and meaning recognition showed more significant gains in the infographics

group. The findings broadly provide further evidence that incorporating visuals with text can lead to improved word learning.

The findings from the four studies on intentional vocabulary learning show that explicit teaching of vocabulary, perhaps including some visuals such as infographics, learning through flashcards and spacing the learning sessions, are all potentially effective tools to improve the limited vocabulary knowledge of Saudi EFL students. There are other ways to improve vocabulary learning intentionally that have been tested with Saudi EFL students including training in vocabulary learning strategies (Alqurashi, 2020) and the use of MALL applications such as Memrise (Almansour, 2019).

Students also need opportunities to learn vocabulary incidentally through reading and viewing. A large body of research has shown that extensive reading can lead to substantial vocabulary learning if sustained over longer periods of time (see section 2.3.1.1). Al-Homoud and Schmitt's study (2009) discussed earlier noted that Saudi EFL students are generally reluctant to read in English, perhaps since reading for pleasure is not very common among the younger generation. Despite this, their study showed that extensive reading can work if students are encouraged to do so. This is supported by a case study in which one Saudi university student was asked to read extensively for eight weeks (Alsaif & Masrai, 2019). Results of pre and posttests showed gains of approximately 540 words (lemmas). These gains are larger than the expected uptake rate from classroom input, which was found to be 160 words according to a previous study that identified an uptake rate of 2.5 words per contact hour. Another case study examined extensive viewing of movies and TV shows with L2 subtitles for nearly 50 hours by a Saudi EFL student in a military academy (Masrai & Milton, 2018b). The aim was to examine the extent of vocabulary and

reading speed development (measured by keyboard strokes) resulting from viewing. After 20 weeks, results of vocabulary and reading tests showed an increase of nearly 900 words (lemmas) and a more than 30% increase in his reading speed. Based on the findings of this study, viewing with L2 subtitles not only leads to vocabulary gains but can also improve the reading abilities of Saudi EFL students. One issue however is that viewing with L2 subtitles requires knowledge of high frequency vocabulary (3000 word-families for 95% coverage; van Zeeland & Schmitt, 2013) which many Saudi school level and some university students fall short of according to the studies in Table 3 above. Not being able to understand the content well might detract from the enjoyment of these activities, leading to frustration and potentially causing individuals to disengage. Therefore, viewing with L1 subtitles might provide more comprehensible and enjoyable input for beginners who have not mastered high frequency vocabulary. Learners can move to L2 subtitles once they develop good knowledge of high frequency vocabulary (Markham et al., 2001). Overall, extensive reading and viewing are important sources of incidental vocabulary learning that can improve the limited vocabulary knowledge of Saudi EFL students.

A balance between intentional and incidental vocabulary learning is important. CLT takes center stage in today's language teaching thinking (Richards, 2006). Due to CLT, with its emphasis on learning language in meaning-focused activities, some teachers and researchers are hesitant to encourage the learning of vocabulary from decontextualized activities such as flashcards and wordlists (Nation, 2011). They worry that this type of rote memorization does not lead to long-term retention or help in word use (Nation, 2022). Folse (2004) notes that this is a common myth in vocabulary instruction that does not receive support from empirical evidence. Research shows that intentional vocabulary

learning usually leads to larger, faster and more enduring gains (Schmitt & Schmitt, 2020). Additionally, both offline (Webb, 2007a) and online (Elgort, 2011) studies found that vocabulary learned from decontextualized activities was not very different in terms of use from vocabulary learned from context. What this suggests is that intentional vocabulary learning is effective and should be one part of a language learning program to maximize vocabulary gains.

The second key part is incidental vocabulary learning which should represent the larger proportion of time allocated for language learning (Nation, 2007). Nation's four strands framework (2007) suggests that a language course should be divided into learning from four strands or components: meaning-focused input (listening and reading), meaning-focused output (speaking and writing), fluency development (i.e., helping learners use their existing language more efficiently) and form-focused instruction (e.g., learning vocabulary from flashcards, studying tenses, improving pronunciation). Each component should roughly be allocated 25% of course time (Nation, 2007). Therefore, while intentional vocabulary learning is effective, it should only take a small part of language learning time (less than 25%). In sum, the low vocabulary knowledge of Saudi EFL students can be addressed more effectively by ensuring that they have opportunities for both incidental and intentional vocabulary learning.

The chapter reviewed research on vocabulary learning, teaching and research in Saudi Arabia. First, it showed that Saudi EFL students have low vocabulary knowledge despite years of formal instruction. Second, the review on vocabulary teaching showed that teachers seem to have some knowledge gaps in how to effectively teach and facilitate vocabulary learning in the classroom. Finally, the review on vocabulary research on Saudi

EFL students highlighted several empirically tested approaches within the Saudi context, alongside general recommendations aimed at enhancing vocabulary learning in Saudi Arabia.

4. Study 1: The vocabulary growth of Saudi EFL learners and the role of individual differences

The general literature review has shown that relatively little is known about L2 vocabulary growth (Dóczi & Kormos, 2015; Pellicer-Sánchez, 2019; Schmitt, 2019; Webb & Nation, 2017)mitt, 2019; Webb & Nation, 2017). This type of research serves key theoretical (e.g., developing models of L2 vocabulary knowledge development; González-Fernández & Schmitt, 2019) and practical purposes (e.g., setting vocabulary learning goals in curricula and textbooks, see section 2.2.4). As shown in section 2.2.4, no previous study appears to have examined vocabulary growth longitudinally (recognition and recall knowledge of form-meaning links) while taking into account the role of individual differences (out-of-class exposure, self-regulation and motivation; Dóczi & Kormos, 2015; Pellicer-Sánchez, 2019). Additionally, the review chapter on vocabulary research in Saudi Arabia (Chapter 3) has also shown limited research on the vocabulary knowledge of school-level students despite being larger in number than university students (Fawaaz, 2023). The current study seeks to examine their vocabulary learning and evaluates the efficacy of current English vocabulary learning and instruction in Saudi Arabia.

The present study aimed to measure receptive vocabulary knowledge development of Saudi EFL students over a school semester (12 weeks). It also examined the effect of out-of-class

exposure, self-regulation and motivation on vocabulary growth (meaning recognition and meaning recall). The study was guided by the following three main research questions:

RQ1. How many words from the most frequent 5000 do learners acquire (to the meaning recognition and meaning recall level) over the course of a semester?

RQ2. What is the role of individual differences, namely: out-of-class exposure, self-regulation and motivation in meaning recognition and meaning recall learning?

RQ3. How do the different components of out-of-class exposure (e.g., watching movies, listening to songs) relate to meaning recognition and meaning recall learning?

4.1 Participants

141 male Saudi EFL students from two educational levels participated in this study: 69 final year intermediate school students and 72 first year secondary school students. The students from each level were recruited from three different classes. All students were enrolled in public schools in Tabuk, Saudi Arabia. Only the students who took week 1 and week 12 tests were included in the study which reduced the number of students from 141 to 103. Of the 103 students, 62 were intermediate students (from three classes, same school: 24, 19 and 19) and 41 were secondary students (from three classes, same school: 17, 17 and 7).

Typically, students finish intermediate school at the age of 14-15 and secondary school at the age of 17-18. Intermediate students receive 3 hours of English instruction per week while secondary students receive 3.75. English language instruction is introduced at grade four (normally at the age of 10) thus the intermediate students have studied English for 5

years while the secondary school students have studied English for 8 years. However, this changed in 2021 and English is now taught to first grade students (this does not affect participants in this study).

The participation in this study was limited to male students since gender segregation laws in Saudi schools prevent males from accessing female schools, which makes recruiting participants and administering relevant instruments difficult. An ethical approval from the University of Birmingham was granted before the study was conducted.

4.2 Instruments

4.2.1 Vocabulary tests

The study used the UVLT to measure meaning recognition knowledge since it is most likely more sensitive to vocabulary growth than other tests (e.g., the VST or CATSS) due to its higher sampling rate of 30 items per 1000 words (see section 2.2.2). The test is also suitable for beginner learners (Nation, 2022), which is the case with the Saudi EFL students here, because it focuses on the first 5000 word families. To measure students' meaning recall knowledge, a meaning recall test was created from the UVLT using the same words (recall UVLT). One advantage of using the same words is to see how recognition and recall knowledge of meaning develop over time since previous research has demonstrated that they are fundamentally different (González-Fernández & Schmitt, 2019) and has also suggested that this distinction might be more significant than the differences between the different vocabulary knowledge components (e.g., form-meaning, derivatives, collocation).

Previous research has shown that bilingual vocabulary tests provide more reliability than monolingual ones, especially with low proficiency learners (Elgort, 2013). Thus, both tests were translated to Arabic and reviewed by two translators with graduate degrees in translation (see Appendix 1 and Appendix 2 for meaning recognition and meaning recall UVLT respectively). Figure 4 shows the format of the Arabic UVLT:

Figure 4. Example recognition items from the Arabic UVLT

	choice	computer	garden	photograph	price	week
سعر					1	
صورة				1		
حديقة			1			

The meaning recall test is essentially a translation test where the L2 form was provided and students had to supply the L1 meaning (see Figure 5).

Figure 5. Example meaning recall items from the Arabic UVLT

Word	Meaning
price	……سعر…….……
photograph	……صورة……
garden	……حديقة…….……

4.2.2 Out-of-class exposure, strategic learning and motivation instruments

Following Peters (2018), the study used part of the European Survey of Language Competences (ESLC; European Commission, 2012) to investigate students' frequency of out-of-class exposure to English. Both the overall exposure score and the scores on the different types of exposure (e.g., watching TV, listening to music) were examined in relation with students' tests scores. The 9 items are shown below in Table 4 with the response categories:

Table 4. Out-of-class exposure items and response categories

Items
1. How often do you listen to songs in English?
2. How often do you watch movies spoken in English without subtitles?
3. How often do you watch movies spoken in English with subtitles?
4. How often do you watch television programs spoken in English without subtitles?
5. How often do you watch television programs spoken in English with subtitles?
6. How often do you play computer games spoken in English?
7. How often do you read books written in English?
8. How often do you read a magazine or a comic written in English?
9. How often do you visit websites written in English?

Reponses				
Never	A few times/year	Once / month	A few times/month	A few times/week

The SRCvoc instrument (Tseng et al., 2006) was used to measure students' self-regulation capacity in learning vocabulary (Appendix 3). As suggested earlier, the instrument is based

on Dörnyei (2001) in which he conceptualizes self-regulation as consisting of five components:

Commitment control	(Maintaining or increasing commitment to achieving goals)
Metacognitive control	(Controlling focus and reducing procrastination)
Satiation control	(Eliminating boredom and making learning more enjoyable)
Emotion control	(Managing emotions by lowering negative feelings and promoting positive ones)
Environmental control	(Harnessing the environment to promote learning)

Each component is measured by four items in the SRCvoc, which add up to 20 items in total. A sample of items include: “Once the novelty of learning vocabulary is gone, I easily become impatient with it” (satiation control), and “When learning vocabulary, I have special techniques to achieve my learning goals” (commitment control). The instrument uses 6-point Likert-scale responses ranging from “strongly agree” to “strongly disagree”. A score of 4 or higher on any item (“slightly agree”) indicates the possibility that a student has control over that dimension. The overall score of self-regulating capacity is obtained by calculating the total score of individual items.

To measure motivation, the study used The Self-Determination Theory of Second Language Scale (SDT-L2; Alamer, 2021a; Appendix 4) given that it has been validated in

the Saudi context and was found to be a reliable measure of L2 motivation (Alamer, 2021b). The SDT-L2 has 20 items with a 5-point Likert-type response format ranging from "strongly agree" to "strongly disagree". The items are structured around the question "Why are you learning English?" with items like "Because I enjoy learning English" (intrinsic orientation) and "Because I want to get better marks in the English course" (external orientation). Unlike self-regulation and out-of-class exposure, the motivation questionnaire does not have one overall score but consists of two main scales (autonomous and controlled motivation) which were used in the analysis. The SDT-L2 (and the other instruments) were translated into Arabic and reviewed by two translators with graduate degrees in translation.

4.3 Procedure

Both the meaning recognition and meaning recall tests were administered in class to the students at the beginning and at the end of a school semester (12 weeks). Each test took place on a separate day to reduce test fatigue. Students took the meaning recall test on day one with the self-regulation questionnaire. On the next day, students took the meaning recognition test with the motivation and out-of-class exposure questionnaires. Meaning recall tests normally take longer time than recognition therefore only one questionnaire was administered with the meaning recall test. They were always given the meaning recall tests first to avoid the possibility of learning from the recognition test options if the order was reversed. To reduce random guessing, students were instructed to skip unknown words yet informed to go through all the words since they might know lower frequency words. Students were given 50 minutes to complete the tests each day although the majority finished within 20-30 minutes. The questionnaires were administered in the first week of

the semester while the vocabulary tests were administered both in week one and week 12 in the third (final) semester of the school year.

4.4 Analysis

The study used generalized linear mixed-effects models (GLMM) to analyze the vocabulary growth of Saudi EFL students and examine the effect of individual differences on the growth estimates (Baayen et al., 2008). GLMM can incorporate both fixed and random effects in one model, which is particularly useful when individual variation between subjects is expected (Linck, 2016). Moreover, when analyzing longitudinal data, GLMM can offer several advantages over other statistical techniques such as repeated-measures ANOVA (Cunnings, 2012; Walker et al., 2019). For instance, mixed-effects model can be used with datasets that have missing data points (almost inevitable in longitudinal studies), unlike other statistical techniques which require the observation data to be dropped resulting in a lower sample size and possibly lower statistical power. Another advantage is the ability to include multiple predictors (fixed effects) in the model at once.

The models in this study were fitted with vocabulary test scores as a dependent variable (scored dichotomously as 1 correct and 0 incorrect). For a small number of cases in the meaning recall tests (mainly unclear handwriting), another rater was consulted, and an inter-rater agreement was reached on all of these cases. Subject (participants) and item (UVLT words) were fitted as random effects while time (week 1 vs. week 12), grade (intermediate vs. secondary), motivation, out-of-class exposure and self-regulation were fixed effects. To ensure an inclusive analysis of the second research question, separate models were constructed for motivation, out-of-class exposure and self-regulation before

combining them in comprehensive models. The models reported have been chosen following a forward selection approach based on likelihood ratio tests. The analysis was conducted through the programming language R (version 4.1.1; Team, 2021) using the lme4 package (version 1.1-27.1; Bates et al., 2015).

High performing students (i.e., outliers) were retained in the analysis since variability in vocabulary knowledge is typical within classroom settings (Dóczi & Kormos, 2015; Kim & Webb, 2022; Webb and Chang, 2012). Additionally, it is difficult to establish objectively an exclusion threshold (Larson-Hall, 2016).

4.5 Results

Table 5 shows the reliability of the instruments used in this study. Item 8 on the autonomous motivation scale was removed to improve the scale's reliability (see Appendix 4). All the instruments had good reliability with Cronbach alpha scores over 0.80. The results of one student on the vocabulary tests were excluded from the analysis due to extreme inconsistency in test scores (scored 5 on the first test and 101 on the second test).

Table 5. Reliability scores of the instruments used in the study

Instrument	Items	Cronbach α
UVLT week 1	150	.96
UVLT week 12	150	.95
Recall UVLT week 1	150	.96
Recall UVLT week 12	150	.98
Autonomous motivation	9	.80
Controlled motivation	10	.81
Self-regulation	20	.81
Out-of-class exposure	9	.81

Note. Autonomous motivation Cronbach α without deleting item 8 = .58

Part 1 of the results section addresses the first research question, which focuses on analyzing the vocabulary growth of students over a school semester. Part 2 covers the second research question which explores the observed variation in students' scores on the tests from the perspective of three individual variation sources: motivation, self-regulation and out-of-class exposure. Part 3 further analyzes students' out-of-class exposure to English with the aim of uncovering which components in particular have significant effects on students' vocabulary growth.

4.5.1 The vocabulary growth of EFL students over a school semester

Table 6 shows students' mean scores on the meaning recognition and meaning recall tests at the beginning and at the end of the semester for those who took both tests. Multiplying

the test score by 33.3 converts the results back to word families (5000 / 150 = 33.33). On average, intermediate students scored 18.68 at the begninning of the semester (equivalent to 622 word-families in total) to the meaning recognition level and finished with a mean score of 27.94 (930 word-families), with an increase of 9.26 (308 word-families). Secondary students, on the other hand, began the semester with a score of 26.41 (879 word-families) and by the end of the semester this number rose to 27.42 (978 word-families) with an increase of 2.97 (99 word-families). A mixed logistic model was fitted with subject and item as random effects and time of test and grade, and their interaction, as fixed effects. Adding by-subject random slopes for time improved model fit significantly. Results showed simple effects of test time (Odds ratio (OR) = 1.94, $z = 4.60$, $p < .001$) and a marginal interaction between test time and grade (OR = 0.67, $z = -1.76$, $p = .078$). Pairwise comparisons using the emmeans package (with Bonferroni adjustment for multiple comparisons) showed that the intermediate group (OR = 1.94, $z = 4.60$, $p < .001$) but not the secondary group (OR = 1.29, $z = 1.43$, $p = .910$) showed an improvement from the first to second meaning recognition tests (see Figure 6).

Students meaning recall knowledge on the other hand is substantially smaller for both groups. Intermediate students mean score was 4.9 (163 word-families) at the beginning and 6.28 (209 word-families) at the end with an increase of 1.38 (46 word-families). Secondary students knew more than twice as much vocabulary compared to intermediate students on both the first test with a mean score of 10.11 (337 word-families) and on the second test with a mean score of 12.57. They also made more than double the gain with a mean score of 2.46 (82 word-families), nevertheless the gains for both groups remain modest. A model similar to the recognition test was fitted to the meaning recall tests data. Results showed

simple effects of test time (OR = 1.57, z = 3.99, p < .001) but no effect of grade (OR = 1.72, z = 0.83, p = .404) and no significant interaction between test time and grade (OR = 1.15, z = 0.88, p = .374). Pairwise comparisons revealed that both the intermediate group (OR = 1.57, z = 3.99, p < .001) and the secondary group (OR = 1.82, z = 5.18, p < .001) showed significant improvement from the first to second meaning recall tests (see Figure 6).

Table 6. Meaning recognition and meaning recall mean scores on week 1 and week 12

		Week 1			Week 12			Growth		
		M	SD	CI	M	SD	CI	M	SD	CI
Recognition	Intermediate	18.68	11.12	[15.52, 21.84]	27.94	12.31	[5.14, 31.44]	9.26	13.11	[5.53, 12.99]
	Secondary	26.41	27.48	[16,82, 36]	29.38	27.42	[19.81, 38.95]	2.97	11.79	[-1.14, 7.08]
Recall	Intermediate	4.9	6.34	[3.1, 6.7]	6.28	6.88	[4.32, 8.24]	1.38	2.94	[0.55, 2.21]
	Secondary	10.11	14.48	[5.14, 15.08]	12.57	18.88	[6.09, 19.05]	2.46	5.74	[0.49, 4.43]

Note. CI = 95% confidence interval

Figure 6. Interactions between grade and test time on meaning recognition and meaning recall tests

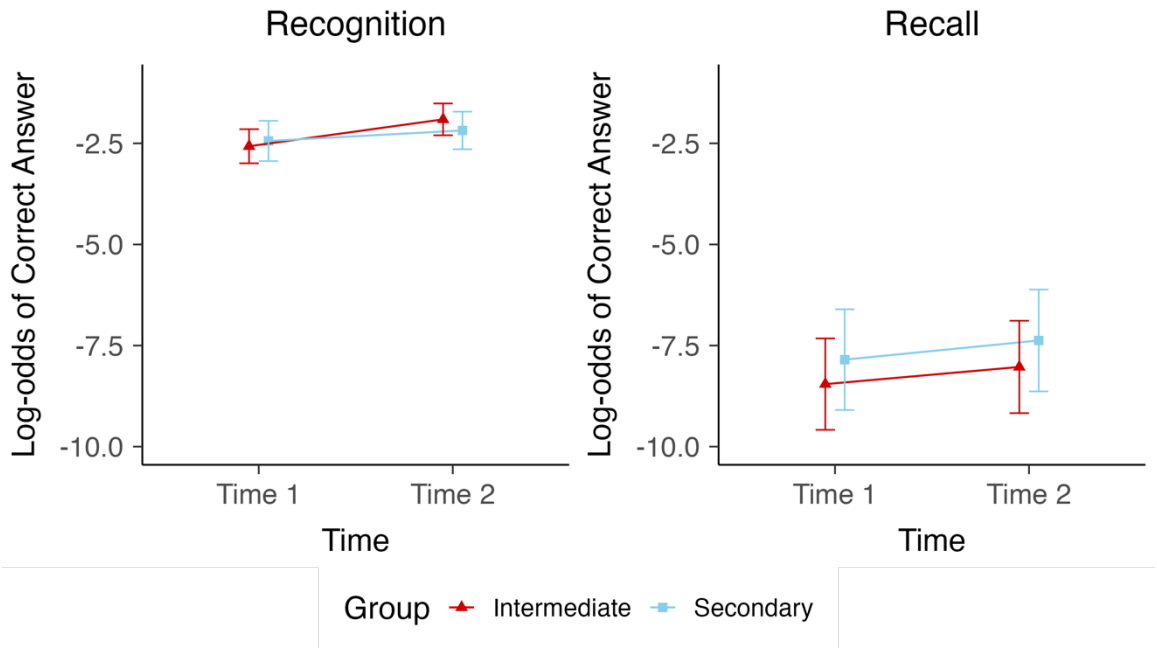
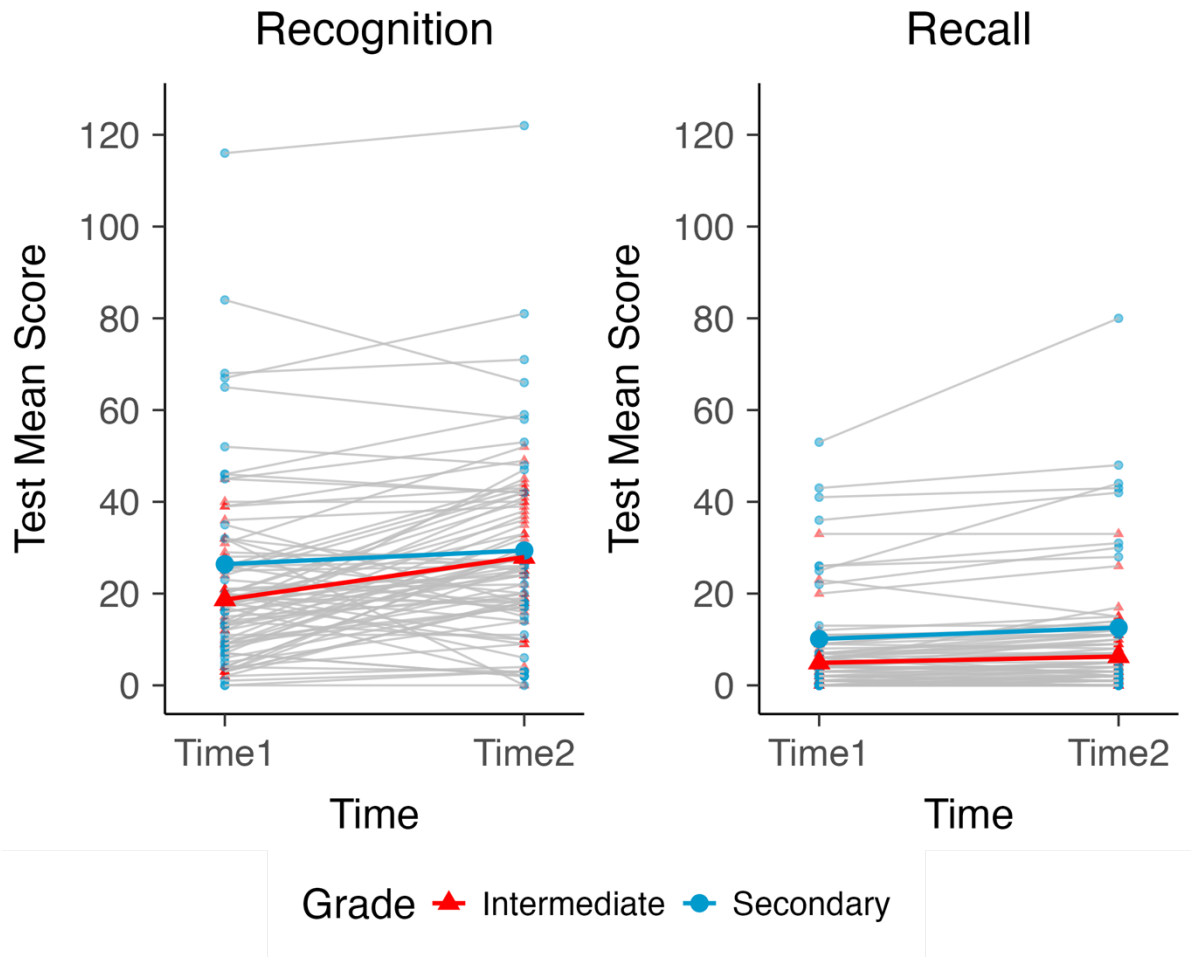


Figure 7 shows the performance of individual students on the first and second tests (meaning recognition and meaning recall). As might be expected, students varied considerably in their starting and ending points. The scores of the majority of students increased by the end of the semester however there are some students whose scores either did not show notable improvement or declined.

Figure 7. Individual growth lines of learners on meaning recognition and meaning recall tests



Note. The red and blue lines represent group mean

Table 7 and Table 8 show students' scores and growth for each frequency band from the meaning recognition and meaning recall tests. As might be expected, students know more frequent words than infrequent words, which can be observed in the general decline of students' scores as word frequency decreases. The authors of the UVLT test recommend a 29/30 threshold for the mastery of the high frequency words (1000, 2000 and 3000) given their importance and 24/30 for the 4000 and 5000 levels. Only one student (0.97%) mastered the first 1000 band on both meaning recognition and meaning recall tests and none mastered the remaining frequency bands.

Table 7. Meaning recognition test mean scores and growth according to frequency band

		1000	2000	3000	4000	5000
Intermediate	Week 1	8.12	3.76	1.98	2.76	2.06
	Week 12	9.76	6.04	4	4.52	3.62
	Growth	1.64	2.28	2.02	1.76	1.56
Secondary	Week 1	10.82	5.68	2.82	4.38	2.71
	Week 12	12.12	7.18	3.09	4.44	2.56
	Growth	1.29	1.50	0.26	0.06	-0.15

Band max score 30; test max score 150

Table 8. Meaning recall test mean scores and growth according to frequency band

		1000	2000	3000	4000	5000
Intermediate	Week 1	3.52	1.04	0	0.22	0.12
	Week 12	4.3	1.32	0.02	0.44	0.2
	Growth	0.78	0.28	0.02	0.22	0.08
Secondary	Week 1	6.09	2.89	0.34	0.49	0.31
	Week 12	7.00	3.26	0.46	0.91	0.94
	Growth	0.91	0.37	0.11	0.43	0.63

On both the first and second meaning recognition tests (Table 7), the highest growth occurred in the 2000 frequency band for both groups whereas the least growth occurred in the 5000 band. In terms of meaning recall knowledge, the highest growth occurred in the first 1000 band whereas the least occurred in the 3000 band (Table 8). Figure 8 and Figure 9 compare meaning

recognition and meaning recall growth of the intermediate and secondary students on each frequency band. For intermediate students (Figure 8), meaning recognition vocabulary growth was larger than meaning recall in all frequency bands. Consistent growth can be seen in all bands for meaning recognition knowledge but not for meaning recall, where there is almost no growth in the 3000 band and a very small growth in the 5000 band. For secondary students (Figure 9), meaning recognition vocabulary growth was larger than meaning recall only in the first three frequency bands while meaning recall growth was larger on the 4000 and 5000 bands. Unlike intermediate students, there is no consistent growth of meaning recognition knowledge across all frequency bands. Overall, meaning recognition growth was generally larger than meaning recall for all students.

Figure 8. Intermediate students' growth on the most frequent 5000 words

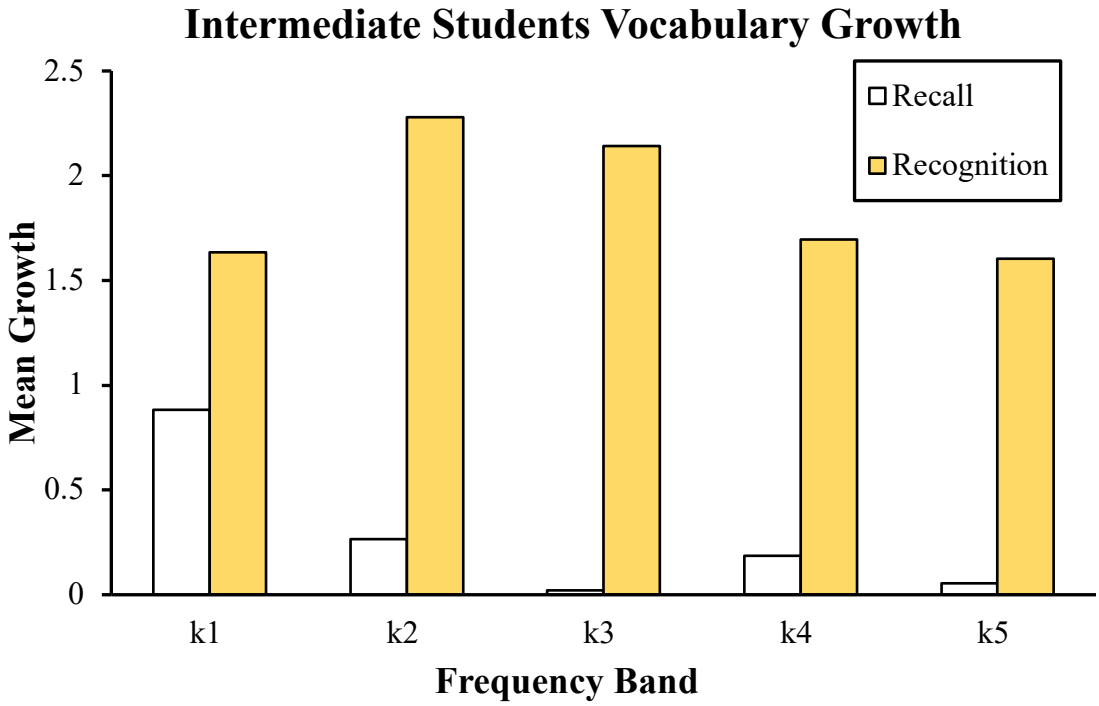
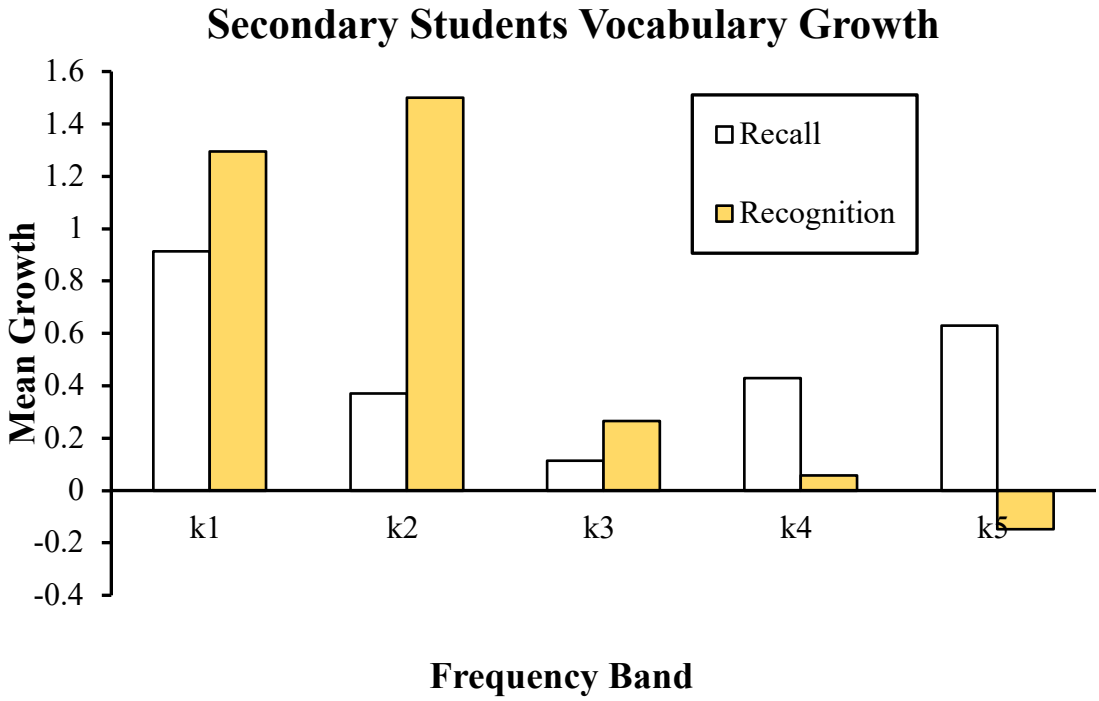


Figure 9. Secondary students' growth on the most frequent 5000 words



4.5.2 The role of motivation, self-regulation and out-of-class exposure in vocabulary learning

Table 9 shows the descriptive statistics for the individual differences. Students in both groups appear to be It shows that intermediate students expressed higher levels of autonomous motivation, controlled motivation and out-of-class exposure than the secondary students. The levels of self-regulation are almost identical.

The results in Table 9 indicate high levels of autonomous motivation, with scores of 4.30 for Intermediate and 4.09 for Secondary learners out of a maximum of 5, showing that learners are primarily motivated by personal interest. Controlled motivation is moderate, with scores of 3.72 (Intermediate) and 3.56 (Secondary), reflecting some influence of external factors. Out-of-class exposure is relatively low for both groups (2.78 and 2.71), suggesting less engagement with the language outside formal learning environments. Self-regulation scores are high (4.59 and 4.60 out of 6), indicating that learners generally manage and control their learning processes effectively.

Table 9. Descriptive statistics of the individual differences by group

	Intermediate		Secondary	
	Mean	SD	Mean	SD
Autonomous Motivation	4.30	0.56	4.09	0.68
Controlled Motivation	3.72	0.72	3.56	0.82
Out-of-class Exposure	2.78	0.87	2.71	0.94
Self-regulation	4.59	0.53	4.60	0.77

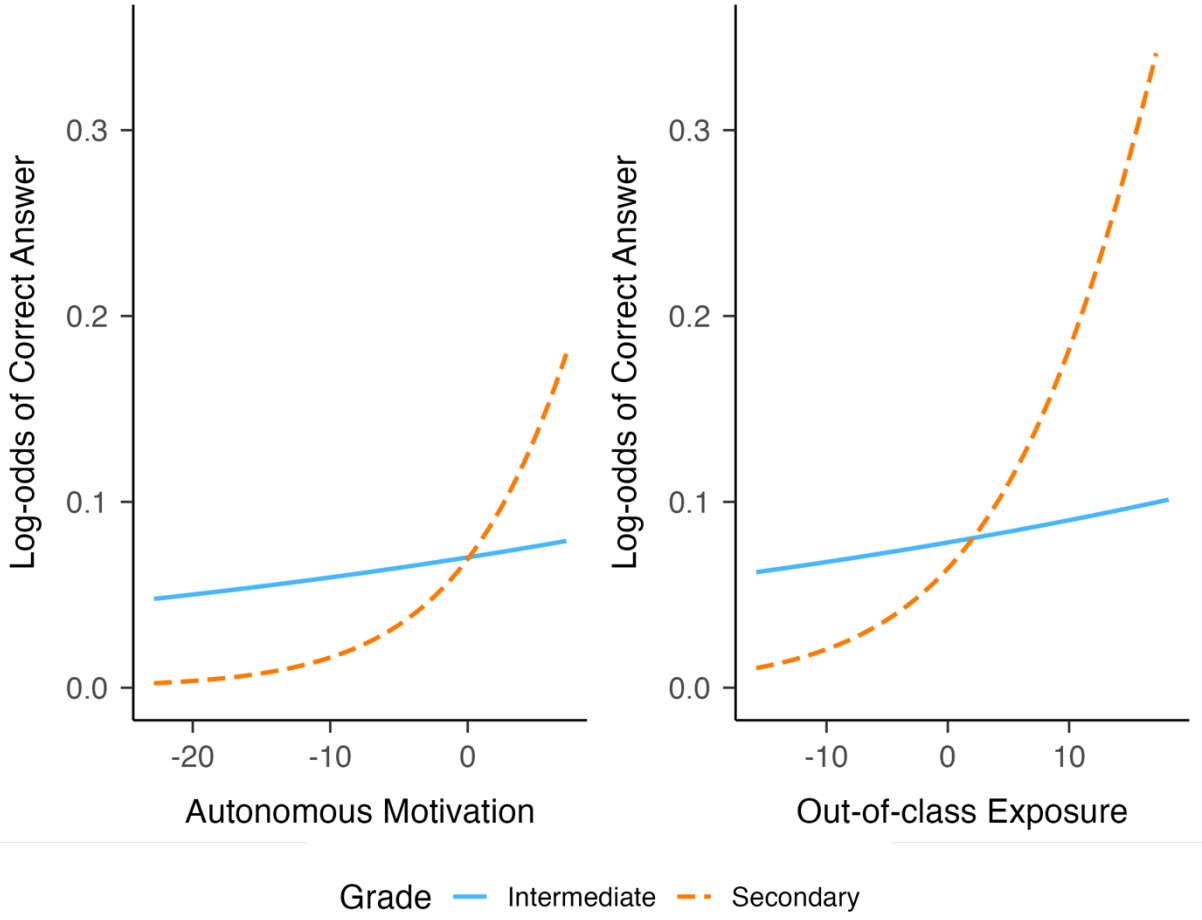
Notes. Max score is 5 for autonomous motivation, controlled motivation and out-of-class exposure. Max score is 6 for self-regulation.

To answer the second research question, separate models were constructed for each of motivation, self-regulation and out-of-class exposure, before combining them in comprehensive models. All of the models had the same base model which consisted of fixed effects of time and grade and an interaction between the two, subject and item as random effects and by-subject random slopes for time. In the separate analysis, all three predictors were fitted in models first without interactions and next in two-way interactions with grade. The separate and comprehensive analyses were conducted once for meaning recognition knowledge and once for meaning recall knowledge.

4.5.2.1 Separate models

Starting with meaning recognition vocabulary knowledge, a model was fitted to examine the effect of autonomous motivation on student test scores. The autonomous motivation model had a significantly better fit compared to the base model ($\chi 2$ (1) = 9.76, p = .001). The model fit improved further with a two-way interaction between motivation and grade ($\chi 2$ (1) = 7.40, p = .006). The simple effect of autonomous motivation on the meaning recognition test scores was not significant (OR = 1.02, z = 0.55, p = .581), however the interaction between autonomous motivation and grade was significant (OR = 1.14, z = 2.76, p = .005). Simple slope analysis (using the interactions package) and visual inspection of the interaction (see Figure 10, left panel) showed that autonomous motivation had a significant positive effect on meaning recognition tests scores for secondary students (OR = 1.16, z = 4.20, p < .001) but not for intermediate (OR = 1.02, z = 0.55, p = .581). Controlled motivation, the second half of motivation in this study, did not have a significant effect on students' meaning recognition tests' scores with or without interactions (all ps >.05).

Figure 10. Interactions plots for autonomous motivation (left) and out-of-class exposure on meaning recognition tests (right)



The effect of out-of-class exposure on students' performance on the meaning recognition tests was examined by fitting a model which had significantly better fit compared to the base model ($\chi 2$ (1) = 20.14, $p < .001$). Adding an interaction between out-of-class exposure and grade further improved model fit significantly ($\chi 2$ (1) = 13.11, $p < .001$). Similar to autonomous motivation, the simple effect of out-of-class exposure was not significant (OR = 1.01, z = 0.80, p = .421) but there was a significant interaction between out-of-class exposure and grade (OR = 1.10, z = 3.72, $p < .001$). Simple slope analysis of the interaction showed that greater levels of out of class exposure significantly improved tests scores for secondary students (OR = 1.13,

z = 6.08, p < .001) but not for intermediate students (OR = 1.03, z = 0.80, p = .422; see Figure 10, right panel).

Self-regulation was fitted in a model which had a marginally signficant better fit than the base model ($\chi 2$ (1) = 3.84, p = .050). Including an interaction with grade did not improve model fit (p >.05). The findings showed a significant main effect of self-regulation on the odds of answering correctly on the meaning recognition tests (OR = 1.01, z = 1.97, p = .048).

In terms of meaning recall vocabulary knowledge, a model was fitted to investigate the effect of autonomous motivation on students' meaning recall tests scores. The model had significantly better fit compared to the base model ($\chi 2$ (1) = 12.71, p < .001), but did not improve significantly with a two-way interaction with grade (p > .05). The results of the model showed that higher levels of autonomous motivation did not have a significant effect on the odds of correct answers (OR = 1.24, z = 3.54, p = .059). Similar to meaning recognition, controlled motivation did not have a significant effect on students' meaning recall tests scores with or without interactions (all ps >.05)

The model for out-of-class exposure on students' meaning recall test scores had signifcantly better fit compared to the base model ($\chi 2$ (1) = 4.74, p = .029). Adding an interaction between out-of-class exposure and grade improved model fit significantly ($\chi 2$ (1) = 7.97, p = .004). The results showed no significant simple effect of out-of-class exposure on meaning recall test scores (OR = 0.97, z = -0.587, p = .557) but showed a significant interaction between out-of-class exposure and secondary grade (OR = 1.23, z = 3.07, p = .002). Results of the simple slope analysis revealed a significant improvement on test scores for secondary students with more frequent out-of-class exposure to English (OR = 1.19, z = 3.82, p < .001) but not for intermediate students (OR = 0.97, z = -0.63, p = .527). A model with self-regulation as a fixed

effect did not improve model fit significantly compared to the base model ($\chi 2$ (1) = 1.02, p = .310) nor did the model with an interaction with grade (p >.05).

4.5.2.2 Comprehensive models

All factors that were shown to be significant in the previous analysis were combined in comprehensive models to examine how they jointly affect meaning recognition and meaning recall vocabulary knowledge. The comprehensive models were then compared to the base models which consisted of time and grade and an interaction between the two as fixed effects, subject and item as random effects and by-subject random slopes for time.

For meaning recognition vocabulary knowledge, the comprehensive model included an interaction between autonomous motivation and grade and another interaction between out-of-class exposure and grade. The model fit was a significant improvement compared to the base model ($\chi 2$ (4) = 38.23, p < .001). The findings show that none of the three individual factors investigated had a significant main effect on meaning recognition tests' scores (Table 10). Only the interaction between out-of-class exposure and grade was significant (OR = 1.08, z = 2.78, p = .005). Simple slope analysis showed that out-of-class exposure to English improved secondary students' scores (OR = 1.10, z = 4.39, p < .001) but not intermediate students' (OR = 1.03, z = 1.56, p = .119). Autonomous motivation did not show significant effect on students meaning recognition tests' scores nor its interaction with grade (all ps >.05).

Table 10. Comprehensive model output for meaning recognition vocabulary growth, motivation and out-of-class exposure

Fixed effects	β	OR	Std. Error	Z value	p
(Intercept)	-2.60	0.07	0.20	-12.74	**< .001**
Week 12 test	0.62	1.86	0.16	3.83	**< .001**
Secondary	-0.10	0.90	0.25	-0.38	0.702
Out-of-class exposure	0.02	1.02	0.02	1.03	0.304
Autonomous motivation	0.02	1.02	0.03	0.80	0.423
Week 12 test * Secondary	-0.03	0.96	0.25	-0.14	0.891
Secondary * Out-of-class exposure	0.08	1.08	0.03	2.78	**0.005**
Secondary * Autonomous motivation	0.05	1.05	0.05	1.12	0.264

Random effects variance (subject = 1.29, item = 1.54)

For meaning recall vocabulary knowledge, the structure of all models is similar to the meaning recognition analysis, with the comprehensive model consisting of autonomous motivation and an interaction between out-of-class exposure and grade. The comprehensive model had better fit compared to the base model ($\chi 2$ (3) = 19.28, $p < .001$). The results of the model (Table 11) show a significant effect for autonomous motivation on meaning recall knowledge. The odds of correct answers on the tests increased for students with higher autonomous motivation (OR = 1.17, z = 2.53, p = .011). The interaction between out-of-class exposure and grade was not significant on meaning recall tests scores (OR = 1.15, z = 1.82, p = .068)

Table 11. Comprehensive model output for recall vocabulary growth, motivation and out-of-class exposure

Fixed effects	β	OR	Std. Error	Z value	p
(Intercept)	-7.93	0.00	0.52	-15.18	**< .001**
Week 12 test	0.38	1.46	0.27	1.39	0.163
Secondary	0.48	1.62	0.59	0.81	0.416
Out-of-class exposure	0.00	1.00	0.05	-0.09	0.926
Autonomous motivation	0.16	1.17	0.06	2.53	**0.012**
Week 12 test * Secondary	0.31	1.37	0.39	0.81	0.419
Secondary * Out-of-class exposure	0.14	1.15	0.08	1.82	0.068

Random effects variance (subject = 6.48, item = 8.98)

4.5.3 An extended analysis of students' out-of-class exposure to English

The out-of-class questionnaire involves several diverse components which might behave differently from one another (e.g., listening to songs is different from gaming in that the latter can involve interaction) whereas motivation and self-regulation components show perhaps more homogeneity (i.e., all components revolve around motivation or self-regulation). Therefore, the aim of this section is to delve into and analyze the nine components of out-of-class exposure activities and examine their relationship with meaning recognition and meaning recall vocabulary knowledge.

Figure 11 shows students' response to the out-of-class exposure questionnaire. The four activities that students engage with most on a weekly basis (most frequent to least) are playing video games, watching movies with L1 subtitles, listening to songs and watching series with L1 subtitles. Conversely, 60% of the students reported that they never read books in English

out-of-class. Finally, viewing movies and series without L1 subtitles is not very common among Saudi EFL students.

Figure 11. Results of the out-of-class exposure questionnaire

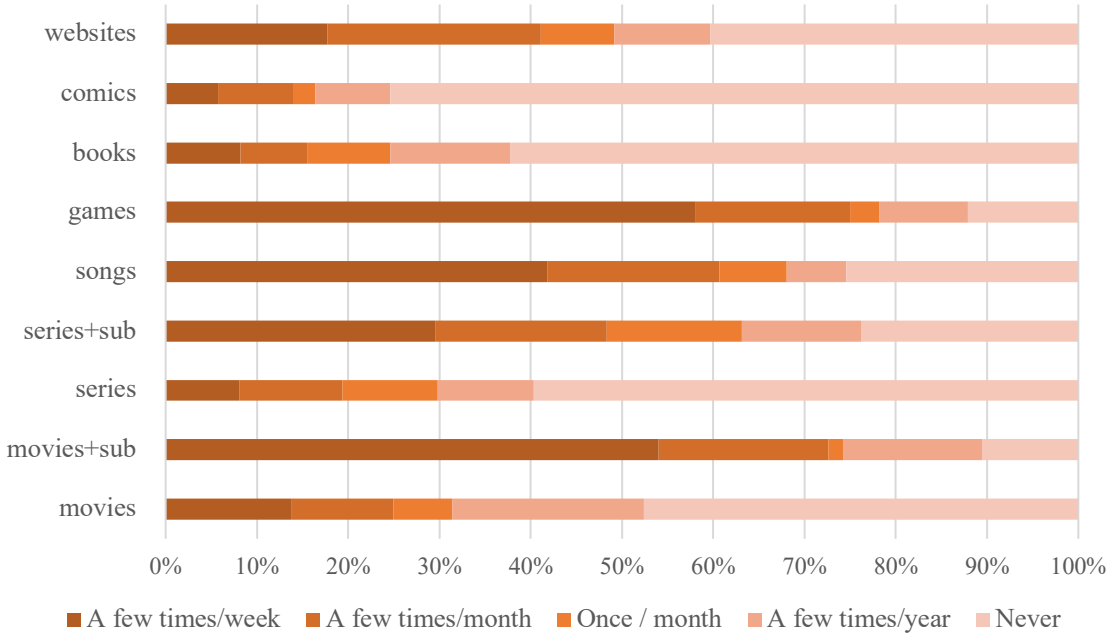


Table 12 shows Spearman's rho correlations between out-of-class exposure components and meaning recognition and meaning recall vocabulary test scores, as the data was not normally distributed. For meaning recognition knowledge, most of the surveyed sources of out-of-class exposure components had significant positive correlation with meaning recognition vocabulary. The only components that correlated significantly with meaning recall test scores were watching movies and series with subtitles, listening to songs, playing video games and visiting websites. They were also the only components that had significant correlation with both meaning recognition and meaning recall vocabulary knowledge.

Table 12. Correlations between the out-of-class exposure components and meaning recognition and meaning recall vocabulary test scores

Input source	Meaning recognition	Meaning recall
Movies	0.09	-0.13
Movies + subtitles	0.23***	0.16*
Series	0.20**	0.10
Series + subtitles	0.27***	0.22**
Songs	0.29***	0.30***
Games	0.30***	0.354***
Books	0.15*	-0.069
Comics	0.10	0.031
Websites	0.28***	0.221**

* < p < .05, ** p < .01, *** p < .001

4.5.3.1 Mixed effects models

Similar to the previous analyses, a model was fitted for each component of the out-of-class exposure questionnaire. Next, the signficant components were combined in comprehensive models that were later compared to the full models with all components to assess their fit. The base model consisted of fixed effects of time and grade and an interaction between the two, subject and item as random effects and by-subject random slopes for time. For meaning recognition vocabulary knowledge, all the components had a significant effect on the meaning recognition tests except for watching movies and series without L1 subtitles and reading books (all ps >.05). A positive effect was found for watching movies (OR = 1.20, z = 2.48, p = .013) and series (OR = 1.24, z = 3.27, p = .001) with L1 subtitles. Similarly, playing video games

(OR = 1.40, z = 4.56, p < .001), listening to songs (OR = 1.32, z = 4.45, p < .001), visiting websites (OR = 1.29, z = 3.86, p < .001) and reading comics (OR = 1.23, z = 2.33, p = .019) all increased students' odds of answering correctly on the meaning recognition tests. The six significant out-of-class exposure components were combined in one comprehensive model. There was no multicollinearity as the variance inflation factor (VIF) for all components was below 2. The full model (with all nine components) did not show significantly improved fit compared to the model with significant components only ($\chi 2$ (3) = 5.38, p = .146). The results of the significant components model (Table 13) show that both playing video games (OR = 1.21, z = 2.48, p = .013) and listening to songs (OR = 1.19, z = 2.26, p = .009) significantly improved students' odds of answering correctly on the meaning recognition tests. Overall, watching English content (movies and series) and reading (books, comics and websites pages) did not significantly increase Saudi students' odds of answering correctly on the meaning recognition tests.

Table 13. Mixed effects output for meaning recognition and out-of-class exposure components

Fixed effects	β	OR	Std. Error	Z value	P
Intercept	-2.61	0.07	0.18	-14.34	**< .001**
Week 12 test	0.68	1.97	0.15	4.63	**< .001**
Secondary	-0.08	0.93	0.22	-0.34	0.731
Movies + subtitles	-0.07	0.94	0.09	-0.76	0.446
Series + subtitles	0.12	1.13	0.08	1.57	0.118
Songs	0.17	1.19	0.07	2.62	**0.009**
Games	0.21	1.24	0.09	2.49	**0.013**
Comics	0.06	1.07	0.09	0.72	0.473
Websites	0.06	1.06	0.08	0.77	0.441
Week 12 test * Secondary	-0.28	0.75	0.23	-1.25	0.211

Random effects variance (subject = 1.31, item = 1.43)

In terms of meaning recall vocabulary knowledge, three of the nine components had significant effects on the tests when assessed individually. A positive effect was found for playing video games (OR = 1.99, z = 3.34, p < .001), listening to songs (OR = 1.70, z = 2.98, p < .002) and visiting websites (OR = 1.48, z = 2.10, p = .035) on the odds of answering correctly on the meaning recall tests. The three significant out-of-class exposure components were combined in one comprehensive model. There was no multicollinearity as the VIF score for all components was below 2. Unlike in meaning recognition knowledge, the full model (with all nine components) did show significantly improved fit compared to the model with significant components only ($\chi 2$ (5) = 17.5, p = .003). The results of the full model (Table 14) show that listening to songs significantly increased the odds of giving a correct answer on the meaning recall tests (OR = 1.55, z = 2.22, p = .027). The odds of answering correctly on the meaning recall tests decreased significantly with the increase of watching movies without L1 subtitles (OR = 0.43, z = -3.50, p < .001). Similar to meaning recognition, watching and reading English content did not have significant positive effects on Saudi students meaning recall tests' scores.

Table 14. Mixed effects output for meaning recall and out-of-class exposure components

Fixed effects	β	OR	Std. Error	Z value	P
Intercept	-7.70	0.00	0.48	-15.95	**< .001**
Week 12 test	0.41	1.51	0.26	1.59	0.112
Secondary	-0.18	0.83	0.56	-0.33	0.741
Movies	-0.85	0.43	0.24	-3.50	**< .001**
Movies + subtitles	-0.27	0.76	0.23	-1.15	0.249
Series	0.35	1.42	0.25	1.37	0.170
Series + subtitles	0.28	1.33	0.21	1.37	0.170
Songs	0.44	1.55	0.20	2.22	**0.027**
Games	0.42	1.53	0.24	1.75	0.079
Books	-0.28	0.75	0.24	-1.19	0.236
Comics	0.08	1.08	0.23	0.34	0.735
Websites	0.23	1.26	0.23	1.00	0.317
Week 12 test * Secondary	0.14	1.15	0.39	0.37	0.712

Random effects variance (subject = 5.77, item = 9.21)

4.6 Discussion

The first research question asked how many words Saudi EFL intermediate and secondary students learn over a school semester to the meaning recognition and meaning recall level. Results showed a significant increase of 9.26 (308 word-families) on the meaning recognition tests for intermediate students and a nonsignificant increase of 2.97 (99 word-families) for secondary students. In terms of meaning recall, results showed a significant, albeit small, increase for both groups with an increase of 1.38 (46 word-families) for intermediate students and 2.46 (82 word-families) for secondary students over the school semester.

Intermediate students' meaning recognition vocabulary growth was three times larger than secondary students. One potential factor has to do with the textbooks used and the number of new words presented at every grade (Alsaif & Milton, 2012). Textbooks at the secondary level in Saudi Arabia were previously found to have less new vocabulary introduced and include more repetition than textbooks from earlier grades (Alsaif & Milton, 2012a). It was found that grade 10 (the level of secondary students in this study) had the least new word-families introduced (fewer than 200) while grade 9 (the intermediate students in this study) had the second largest number of new word-families presented (around 900). In fact, the average number of word families introduced at a single intermediate year was larger than the total number of word-families introduced at all three years of secondary education combined (790). Alsaif and Milton (2012) suggest that diminishing the numbers of new words introduced in secondary education seems unwise and might run the risk of halting the learning progress of secondary students. The findings of the current study give further support to their position as the secondary students showed no significant growth over the school semester on meaning recognition knowledge. Although the textbooks might have changed throughout the years, based on the findings here, the same strategy (reducing the amount of new vocabulary introduced) seems to be followed today.

Intermediate students' meaning recognition vocabulary gains over a single school semester seem relatively large when compared to the reported 400 word-families annual gain of other EFL learners (Webb & Chang, 2012). This might give an indication that intermediate students are performing very well when it comes to vocabulary learning. However, a more representative picture of Saudi EFL students' vocabulary knowledge can be obtained by examining their scores on the frequency bands and their overall test score. The results revealed clear deficiencies in mastering the highest frequency words after more than six years of instruction. Only one student out of 103 (0.97%) mastered words from the first 1000 frequency

band, and none mastered the remaining bands. Similar findings emerge when students' overall score is examined. The findings suggest that Saudi EFL learners finish intermediate school with an overall meaning recognition score of around 930 word-families and an overall meaning recall score of around 209 word-families. On the other hand, secondary students finish the first year with an overall meaning recognition score of around 979 word-families and an overall meaning recall score of around 418 word-families. Both groups have a meaning recognition overall score below 1000 and a meaning recall overall score under 500. The meaning recognition vocabulary scores found in this study are very close to the 1000 word-families figure reported in two of the three previous studies conducted with secondary Saudi students from different cities[10] (see Table 3 Alhaj et al., 2019; Al-Hazemi, 1993; see Table 3). Research suggests that 2000 to 3000 word-families is a necessary component to understand 95% of daily conversations, movies and newspapers (van Zeeland & Schmitt, 2013). Based on their current overall vocabulary scores, Saudi intermediate and secondary students are likely to face serious comprehension difficulties when reading and listening to unmodified or ungraded input. To summarize, although intermediate students had relatively good growth over three months of language learning, their knowledge of high frequency words and overall vocabulary (and that of secondary students) remained modest.

There are several reasons behind the low vocabulary knowledge of Saudi EFL students, one of which might be the lack of exposure to large amounts of input in English: a situation commonly found in many EFL contexts (Siyanova-Chanturia & Webb, 2016). For example, Table 9 has shown relatively low exposure to English out-of-class with most students engaging once a month with English language activities such as watching movies and reading English books.

[10] Alhaj et al. (2019) was conducted in Abha (south) while Al-Hazemi (1993) was conducted in Riyadh, the capital of Saudi Arabia (center). The current study was conducted in Tabuk, in the north of Saudi Arabia.

In terms of school instruction, the current 3 – 3.45 hours of weekly English instruction Saudi school-level students receive do not seem to be sufficient to help them learn even the highest frequency vocabulary (first 1000 word families) after many years of instruction. Another factor could be related to teaching and how vocabulary is handled in class. School teachers in Saudi Arabia seem to lack awareness in how to develop students' vocabulary (Altalhab, 2014; Sonbul et al., 2022). For example, it has been noted that school teachers in Saudi Arabia often pay no attention to teaching unknown words in context (Sonbul et al., 2022). These factors are obviously not exhaustive and other factors such as learning in overcrowded classes play a role in the low vocabulary and language knowledge of Saudi EFL students (Alrabai, 2016).

Table 6 shows relatively large SD values. Studies indicate that classrooms commonly exhibit wide vocabulary variability due to differing levels of exposure and due to individual differences among learners (e.g., Dóczi & Kormos, 2015; Kim & Webb, 2022; Webb and Chang, 2012). For example, Webb and Chang (2012) found that one group of participants acquired as few as 18 words over the course of a year, while another group learned up to 430 word-families. Therefore, it's normal for some students to possess larger vocabularies, resulting in high standard deviations

The second research question investigated the role of out-of-class exposure, self-regulation and motivation in vocabulary learning. Results of comprehensive models showed that students with higher autonomous motivation learned more words to the meaning recall level regardless of grade. These findings confirm the important role of autonomous motivation in vocabulary knowledge development as students with higher levels of autonomous motivation are more likely to pursue opportunities to learn language and expand their vocabulary knowledge (Alamer, 2021a; Y. Zhang et al., 2017). Meanwhile, no significant effect was found for autonomous motivation on meaning recognition vocabulary knowledge. One possible explanation is that given that recalling a word from memory is more challenging compared to

recognizing it (González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004), the impact of autonomous motivation might be more noticeable in the act of recalling compared to meaning recognition. In other words, more autonomously motivated learners seem to be better distinguished from less motivated learners when the task is more challenging. In a study examining the relationship between motivation and perceived task complexity, Kyndt et al. (2011, p. 146) found that " [a]pparently, motivation for learning, more specifically autonomous motivation, is only significant or important when students are placed in a context that is designed to have a high workload". Additionally, research from cognitive psychology, motivational and social psychology and the neurosciences provide evidence that motivation varies dynamically with task complexity (Jurczyk et al., 2019).

Controlled motivation did not have a significant influence on students' scores either on meaning recall or on meaning recognition vocabulary. The finding is not surprising given that previous studies mostly found negative or no effect of controlled motivation on language learning outcomes (Alamer, 2021a; Liu, 2007; Noels et al., 1999; F. X. Wang, 2008). Students driven by controlled motivation find less pleasure and interest in language learning compared to students with autonomous motivation therefore they are less likely to seek opportunities to learn language in-class and out-of-class (Noels et al., 2019).

Few studies have examined the role of self-regulation in vocabulary learning, despite this being introduced to vocabulary research more than 15 years ago (Rose et al., 2018). Given the incremental and lengthy nature of vocabulary learning where students are expected to learn thousands of words, self-regulation skills such as planning and monitoring vocabulary learning become very relevant. This is supported by findings showing significant correlations between self-regulation and language learning outcomes (Seker, 2016). The present results, however, showed no significant effect for self-regulation on vocabulary learning in the comprehensive models. One possibility could be that the instrument used needed some adaptation before being

used in different contexts. For example, Mizumoto & Takeuchi (2011) adapted the SRCvoc instrument to the Japanese context and found that the factor structures were different from the original study. Therefore, adapting the instrument to the Saudi context through construct validation using EFA or confirmatory factor analysis and making adjustments accordingly might show more of an effect of this factor (see Alamer et al., 2024).

Results revealed higher gains for secondary students with more frequent out-of-class exposure on meaning recognition vocabulary knowledge. The results are in line with previous research which highlights the positive effect of out-of-class exposure on learners' vocabulary knowledge (De Wilde et al., 2019; Peters, 2018; Sundqvist, 2009). When all components of out-of-class exposure were combined in comprehensive models, listening to songs and playing video games emerged as significant predictors of vocabulary learning. Listening to songs in particular was the only source of input that had a significant effect on both meaning recognition and meaning recall vocabulary. Pavia et al. (2019) found a significant effect for listening to songs on both recognition and recall spoken vocabulary knowledge. They list a number of features that make listening to songs an important source of vocabulary learning such as the repetitive nature of this activity. Repetition in turn is a key condition in promoting vocabulary learning (Pellicer-Sánchez, 2016; Uchihara et al., 2019; Webb, 2007b). In contrast, watching movies without subtitles was significantly associated with lower test scores. It is generally not recommended for beginners with vocabulary knowledge below 2000 word-families (95% coverage) to watch movies without subtitles for learning given the overwhelming number of unknown words they need to decode to understand the content (Webb & Rodgers, 2009a). It seems that the low proficiency students followed the unproductive strategy of watching movies without subtitles whereas the more proficient students appear to have avoided it.

Intermediate students' out-of-class exposure did not have a significant effect on their test scores. This might be related to intermediate students having exposure to more words in-class

than secondary students, possibly making out-of-class exposure less influential. Secondary students on the other hand have fewer words available to them, therefore out-of-class exposure seems to be a significant source of vocabulary learning. Since seeking language learning opportunities out-of-class requires more motivation (Bailey, 2011), this may explain why motivation had a significant effect for secondary students but not for intermediate students.

Individual differences also can explain the scores of the highest scoring students in the study (see Figure 7). A closer inspection of his response to the questionnaires shows that in comparison to the secondary group average (see Table 3), he reported higher levels of autonomous motivation (4.30), expressed lower levels of controlled motivation (2.50), higher levels of out-of-class exposure (3.78) and self-regulation (3.70). These findings highlight the key role of individual differences in explaining variation in vocabulary achievement.

Despite cultural differences, young language learners today might not be very different in their out-of-class exposure to English. A comparison between Saudi and Flemish learners of English (Peters, 2018a) shows a number of similarities. For example, the most frequent activities in both groups included watching movies and shows, listening to songs and playing video games, while the least frequent activities were reading books and magazines. The fact that extensive reading is not very common among Saudi and Flemish learners makes us question its appeal to learners and in turn its effectiveness. Even if there exists large evidence supporting the effectiveness of extensive reading in controlled experiments (Nation & Waring, 2020), its effectiveness is inherently linked to (and limited by) students' interest in doing it. If learners in other contexts also do not voluntarily engage with extensive reading then this may warrant a reconsideration of its perhaps overemphasized status in SLA. One example of this overstatement is Nation's (2022, p. 590) suggestion that "[t]he single most effective improvement that a teacher could make to a course on learning English as a foreign language is to include an extensive reading program". A key message perhaps here is that researchers

and instructors should take into consideration students' preferences when recommending language learning activities (especially out-of-class) and not rely solely on which activity is more effective on paper.

4.7 Pedagogical implications

Vocabulary researchers over the years have provided practical recommendations to improve learners' vocabulary knowledge (Barclay & Schmitt, 2019; Nation, 2020; Schmitt, 2008; Siyanova-Chanturia & Webb, 2016; Webb & Nation, 2017). The findings of this study highlight three key pedagogical implications. Firstly, an intentional vocabulary learning component should be included in EFL language learning programs to assist learners master the highest frequency words. This can take several forms, the most effective of which are flashcards and wordlists (Webb et al., 2020). Secondly, learners should be encouraged to increase their exposure to English out-of-class through activities they prefer such as playing video games, watching movies and listening to songs which have the potential to promote their incidental vocabulary learning. Lastly, for learners to engage effectively in intentional and incidental vocabulary learning (especially out-of-class), they need to be intrinsically motivated. Jones et al. (2009) suggest a number of activities that can be applied in foreign language classrooms to promote learners' intrinsic motivation. What might prove fruitful in terms of increasing learners' motivation in learning the highest frequency words is breaking the words down into a smaller more manageable number by focusing, for example, on learning the Essential Word List (EWL; Dang & Webb, 2016) in the early stages. Similarly, informing learners that the EWL represent 75% of English might help in improving their motivation and make the effort of learning these words seem more worthwhile.

4.8 Conclusion

The vocabulary growth of EFL learners globally is low and slow (Siyanova-Chanturia & Webb, 2016). Despite intermediate students making relatively good growth over the semester, they (along with the secondary students) fell short of mastering the highest frequency words and still had low overall meaning recognition and meaning recall vocabulary knowledge after more than six years of instruction. Variation in vocabulary learning can be partially accounted for by examining the individual differences among learners. Students who had more out-of-class exposure to English and higher motivation generally learned more words by the end of the semester. The findings from the present longitudinal study support the significant role of these factors in vocabulary knowledge development. A closer look at students' out-of-class exposure to English showed that the most frequent activities students engage with weekly are playing video games, watching subtitled movies and series and listening to songs. On the other hand, the most infrequent activities are reading books, comics and watching unsubtitled movies and series. Although extensive reading is considered one of the most effective sources of language learning, 60% of the students reported that they never read English books or comics out-of-class. Developing EFL learners' knowledge of the highest frequency words is a priority in any language learning program. This needs to be developed through intentional and incidental learning in-class and out-of-class within an intrinsically motivating environment.

5. Study 2: Encouraging out-of-class vocabulary learning from digital flashcards through frequent quizzes

The second study was planned to be an intervention study with the objective of improving Saudi students' knowledge of the most frequent words. Results of the first study clearly showed that intermediate and secondary Saudi EFL students fell short of mastering even the first 1000 frequency words. This means that Saudi EFL students will face serious difficulties even understanding daily spoken language (2000-3000 word-families are required for 95% coverage) (van Zeeland & Schmitt, 2013) and authentic written language (Nation, 2006). The goal of the second study was to address this issue through intentional vocabulary learning.

5.1 Background

Research on vocabulary has consistently shown that intentional learning of vocabulary results in more gains compared to incidental learning (Hulstijn, 2001; Laufer, 2003, 2005; see section 3.1.5). Given the fact that high frequency vocabulary (most frequent 3000 word-families) represents the majority of language typically encountered by learners (around 93% coverage; Schmitt & Schmitt, 2014; see section 3.1.2), intentional learning of these words is clearly worthwhile. Intentional vocabulary learning can take several forms including learning from flashcards, wordlists and form-meaning matching (Morgan & Rinvolucri, 2004; Webb & Nation, 2017). A meta-analysis by Webb et al., (2020) examined the effectiveness of intentional vocabulary learning in developing form recall and meaning recall knowledge. Results revealed a 60% gain on meaning-recall test and a 58% gain on form-recall test. The gains however dropped to 39% and 25% on the delayed posttests. Despite the decrease on the

delayed posttests, the gains remain larger than the 9-18% gain from incidental learning reported in another meta-analysis (Webb et al., 2023). The study by Webb et al. (2020) also found that the most effective intentional vocabulary learning activities (as measured by effect size) were flashcards and wordlists. Studies comparing the two show that flashcard use generally results in more gains and it is preferred by learners (Yüksel et al., 2020; Zakian et al., 2022).

5.1.1 Flashcard vocabulary learning

Learning from flashcards typically entails creating connections between L2 words and their meanings (Nation, 2022). The meanings can take several forms including L2 definition, L1 translation or pictures. What makes vocabulary learning from flashcards perhaps the most efficient technique in terms of retention is retrieval (Nation, 2022). Simply put, receptive retrieval in vocabulary occurs when a word is present without its meaning which the learner has to recall from memory while productive retrieval involves retrieving word form. This is perhaps why learning from flashcards is more effective than learning from wordlists (Yüksel et al., 2020; Zakian et al., 2022), since in wordlists the word and its meaning are presented together at the same time while in flashcards only one is available at a time (Nation, 2022). This simple technique has been supported by extensive research over the past decades (e.g., Barcroft, 2007; Karpicke & Roediger, 2008).

Learning from flashcards involves several small yet important decisions to be made to increase its effectiveness (for a review see Nakata, 2020; Nation, 2022). Few of these are straightforward and the majority require careful examination depending on the context of learning. The more straightforward recommendations include, for instance, spacing the learning, which is more effective than massed learning for long-term retention (Baddeley, 1990; Kornell, 2009). For example, learning a group of words over several days is more beneficial for long-term retention than learning them in one day. Another recommendation is

changing the order of the cards frequently to avoid serial learning where one word assists the recall of the following one (Nation, 2022). Other decisions such as choosing the number of words to be learned during a single study session and the direction of learning (receptive where meaning is recalled or productive where form is recalled) are debated and perhaps no one-size-fits-all recommendations can be made. For example, Nakata (2016, 2020) recommends learning productively as receptive knowledge develops along the way in this direction but not the other way around while Nation (2022) recommends learning receptively first, arguing that receptive learning is usually easier for beginner learners. Both, however, agree that researchers and teachers should consider their context-specific factors (e.g., learners' level, whether productive knowledge is a priority and time available for learning) when making choices regarding how flashcard learning is implemented.

Digital flashcards perform the same function as paper flashcards but offer more options that might potentially make them more effective (for a review of flashcard apps, see Nakata, 2011). From a theoretical perspective, digital flashcards, especially in the form of a smartphone app, offer a number of advantages over traditional paper flashcards for learners such as automatic adaptive sequencing (where words that each learner finds more difficult are repeated more frequently), engagement (through gamification elements such as sounds and games), mobility (having the flashcards in an app is more convenient than carrying around physical ones) and assistance in remembering to learn (through notifications). However, empirical evidence does not conform precisely to the suggested theoretical advantages as studies comparing digital and paper flashcards show conflicting findings. Some studies have found an advantage for digital flashcards (Ashcroft et al., 2018; Xodabande, Asadi, et al., 2022; Xodabande, Iravi, et al., 2022; Xodabande, Pourhassan, et al., 2022), citing factors such as multimedia features (audio for pronunciation and picture definitions) as potential reasons, while others found no such advantage (Dizon & Tang, 2017; Nikoopour & Kazemi, 2014; Sage et al., 2019, 2020). The

conflicting findings seem to warrant a comprehensive synthesis such as a meta-analysis to investigate the advantages (or lack of) of digital flashcards over paper flashcards. Until such research is conducted, it is perhaps preferable to use digital flashcards in out-of-class settings if it is accessible to students. This is due to the range of advantages discussed earlier and other teacher-related advantages such as tracking the progress of every learner, the ease of creating the flashcards and the lower costs associated with digital flashcards.

In terms of out-of-class vocabulary learning, several studies showed large and significant gains from flashcard learning (McLean et al., 2013; Xodabande, Pourhassan, et al., 2022; Zakian et al., 2022). One study by McLean et al. (2013) examined out-of-class vocabulary learning from digital flashcards over an academic year as measured by the Vocabulary Size Test (Nation & Beglar, 2007). Group one was required to spend two hours a week learning words from a flashcard website using their computers, the second (control) group was instructed to spend two hours a week on extensive reading and the third group did one hour of flashcard learning and one hour of extensive reading. Results showed growth of 1107 word-families for the flashcard only group, 75 word-families for the extensive reading group and 1147 for the flashcard plus extensive reading group. The findings show that the vocabulary gains from flashcard learning were ten times larger than extensive reading gains.

Taken together, the findings from previous studies suggest that flashcards can be an effective intervention, and their deployment with low proficiency learners such as Saudi EFL students might help in improving their knowledge of high-frequency English words. However, one issue with including flashcards as an out-of-class learning activity is that learners might not be motivated enough to participate. A study by Platzer (2020) showed that 35% of the students did not access an online flashcards platform even once out-of-class. The number rises to 74% if out-of-class flashcard learning is not a mandatory course requirement (Seibert Hanson &

Brown, 2019). One way to encourage students to learn from flashcards in out-of-class settings is through frequent vocabulary quizzes.

5.1.2 Testing effect

Although testing is commonly seen as an assessment tool, research suggests that it can function as a learning tool too (Kanayama & Kasahara, 2018; Karpicke & Roediger, 2007, 2008). This is referred to as the testing effect and it is commonly divided into direct and indirect effects (Roediger & Karpicke, 2006). The direct effects of testing (commonly referred to as just the testing effect) describe the phenomenon in which taking a test on previously learned information leads to better retention than relearning (Yang et al., 2021). One of the studies that examined the testing effect in foreign language vocabulary learning is Roediger and Karpicke (2008), who asked English speaking college students to learn 40 Swahili words using digital flashcards in experimental settings. Students were assigned to one of three treatment groups. Once a student was able to answer a word correctly, the word was either dropped from further testing, dropped from further learning or dropped from further testing and learning. The results of a one-week delayed posttest showed that dropping words from further testing led to significantly lower long-term retention (80% to 36%) compared to dropping words from further learning (80% to 80%). That is, retention is increased if students continue to take quizzes frequently compared to frequently restudying. The findings suggest that testing is not merely a neutral assessment activity but can be a powerful learning tool that enhances the long-term retention of information.

The indirect effects of testing refer to the other test-related effects such as the increased studying time resulting from frequent quizzing (Michael, 1991), the identification of gaps in knowledge (Amlund et al., 1986) and improved knowledge organization (Zaromb & Roediger, 2010). More relevant to out-of-class vocabulary learning is the fact that testing paired with

class credits can function as an incentive to learn and perhaps increase study time (Michael, 1991). Tuckman states:

> "in anticipation of a situation in which a person is required to perform, that person may expend considerable effort in preparation because of the mediation provided by the desire to achieve success or avoid failure. That desire would be said to provide incentive motivation for the person to expend the effort" (Tuckman, 1998, p. 142).

This idea, that testing encourages more studying as measured by a test score, receives support from a meta-analysis study which showed that taking at least one test over a 15-week semester led to one half standard deviation increase on test scores compared to having no tests (Bangert-Drowns et al., 1991). Other studies have also found that testing encourages students to prepare better before class by reading assigned material (Weinstein et al., 2014), increases study time (Yang et al., 2017) and increases attendance (Schrank, 2016).

Overall, testing (with both its direct and indirect effects) seems to be an effective tool to improve students' engagement with out-of-class vocabulary learning. This raises an important question which is how frequent these tests should be (e.g., weekly, biweekly or monthly) and what the effect is of different frequencies of quizzes on vocabulary learning.

5.1.3 Quiz frequency

Experimental psychological research on different quiz frequencies has been around since the early years of the 20th century (Beaulieu & Zar, 1986; Dustin, 1971; Keys, 1934; Palmer, 1974; Ross & Henry, 1939). One of the earliest studies was conducted by Keys (1934), who compared the effects of weekly and monthly quizzes on true and false subject-matter statements over a school semester. The results of his study showed that the weekly group scored 12% higher than the monthly group on final examination. On the contrary, Beaulieu and Zar (1986) found no advantage for weekly quizzes over monthly quizzes on a comprehensive test at the end of the

semester. In their study, 100 college students were assigned to either weekly (12 quizzes) or monthly quiz groups (3 quizzes) with both taking a comprehensive test at the end of the semester. There was a significant advantage for the weekly group on the first round of testing (the monthly group first test mean was 73.02; the weekly group mean score of the first six quizzes combined was 77.14). However, this advantage disappeared by the end of the semester on the comprehensive test (weekly mean score = 71.88; monthly mean score = 72.83). The authors suggest that students might have found the weekly quizzes tedious and possibly exerted less effort as the semester progressed. They postulated that weekly quizzes might have been too frequent and less frequent quizzes might be more optimal.

Bangert-Drowns' (1991) meta-analysis on the frequency of testing examined 35 studies with a minimum testing frequency of zero to a maximum of 48 tests during a semester. 29 studies found positive effects on criterion exams while six studies found negative effects. Of the 29 studies with positive effects, 13 were statistically significant while only one from the negative effects was significant. In terms of the effect size, students who at least had one interim test during a semester scored one half standard deviation higher on a final test than students who had no interim test. This increase however comes with a diminishing return caveat in the sense that increasing the frequency of testing from one to two per semester leads to a small increase of 0.08 standard deviation. The findings show that having tests is effective but having frequent tests might not largely increase the testing effectiveness.

The research on the frequency of quizzes is clear in that having quizzes results in more learning compared to no quizzes. The research however is less clear on the amount of additional learning resulting from the increased frequency of quizzes.

5.1.4 Students' perceptions of frequent quizzes and digital flashcard learning

Contrary to popular belief, students across a number of studies have expressed positive perceptions of more frequent quizzes (Bangert-Drowns et al., 1991; Deck, 1998; Gokcora & DePaulo, 2018; Kika et al., 1992). Students in some studies reported that frequent quizzes helped them enjoy class more (Bangert-Drowns et al., 1991) and increased their confidence on final achievement tests (Gokcora & DePaulo, 2018). Students also appear to have a positive perception towards digital flashcard learning (Davie & Hilber, 2015; Sage et al., 2019, 2020) and mobile-assisted language learning (MALL) in general (Shadiev et al., 2017). By contrast, vocabulary learning from flashcards was described as a "bitter pill" in one study given that it was effective but not enjoyed by students (Seibert Hanson & Brown, 2019). Apart from a few exceptions, students in general seem to have a positive perceptions of frequent quizzes and digital flashcard learning.

One of the studies that have examined the relationship between perceptions or attitude and language learning is Mantle-Bromley (1995). She conducted a study with middle-school students in a short 9-week foreign languages program. The study aimed at maintaining/improving students' attitudes and perceptions of French and Spanish speakers. They were divided into a treatment and a control group. The treatment group received culture-related lessons designed using attitude-change theory. Results showed that the treatment group had significantly better perceptions and attitudes, as measured by questionnaires. Additionally, the study explored students' beliefs about language learning and found that many students had misconceptions that could impede their development in language learning.

In terms of vocabulary, previous research suggests that learners' perceptions regarding vocabulary learning from digital flashcards seems to be related to vocabulary learning (Sage et al., 2019, 2020; Seibert Hanson & Brown, 2019). For example, one of the research questions

in Sage et al.'s (2019) study examined the relationship between recall vocabulary knowledge and students' perceptions of the vocabulary learning medium (paper vs. digital flashcards). The perceptions questionnaire in their study focused on learning satisfaction (extremely dissatisfied to extremely satisfied), perceived control (e.g., "How in control of the flashcards did you feel?") and perceived difficulty (e.g., "How difficult was it for you to learn these words from flashcards?"). The researchers found that vocabulary recall was significantly correlated with perceptions. More specifically, recall knowledge correlated significantly with satisfaction ($r = .48$), control ($r = .28$) and difficulty ($r = .64$).

The previous discussion suggests that learners' perceptions of frequent quizzes and digital flashcard learning are generally positive. It also suggests that perceptions of the learning situation (e.g., quizzes and digital flashcard learning) correlates with language and vocabulary learning outcomes.

5.2 The present study

The previous discussion of the literature points out two key areas that have received little attention. First, it is unclear based on the available research how frequently quizzes should occur for optimal out-of-class vocabulary learning. Second, the majority of research on vocabulary learning from flashcards was conducted in-class, making it difficult to generalize the findings to out-of-class settings. This is unfortunate given that flashcard learning lends itself to being an out-of-class autonomously performed task. One advantage of having flashcard learning as an out-of-class activity lies in the fact that classroom time is usually limited in foreign language instruction (Lightbown & Spada, 2020) and it perhaps should be reserved for tasks that require assistance from the teacher or collaboration with other learners (see Richards, 2015). Therefore, the present study aimed to address these gaps in the literature by examining

the effect of quiz frequency on vocabulary learning from digital flashcards in out-of-class settings.

5.2.1 Method

The current study aimed to increase Saudi EFL students' knowledge of high frequency words through the use of digital flashcards in out-of-class settings. It investigated the role of quiz frequency and individual differences in out-of-class vocabulary learning from digital flashcards. The five research questions addressed were as follows:

> RQ1. Do frequent vocabulary quizzes (in any frequency) lead to globally more vocabulary learning compared to no quizzes?
>
> RQ2. Which of the three types of quizzes frequency (weekly, biweekly, monthly) result in more vocabulary learning by the end of the semester as measured by a posttest?
>
> RQ3. What role do individual differences, namely motivation and self-regulation, play in out-of-class vocabulary learning from digital flashcards?
>
> RQ4. Does more frequent quizzing lead to more studying as measured by the total number of minutes students spend learning from digital flashcards? Does study time affect vocabulary learning?
>
> RQ5. What are students' perceptions of frequent quizzing and digital flashcards learning and how does their perceptions affect vocabulary learning?

5.2.2 Participants

First year secondary school male students aged 16-17 years from four classes participated in the study. The total number of students was 105 on meaning recall tests (weekly = 25, biweekly = 28, monthly = 26, no-quiz= 26) and 101 on meaning recognition (weekly = 25, biweekly = 26, monthly = 25, no-quiz = 25). The number of students who took both the pretest and posttest was 76 on meaning recall (weekly = 20, biweekly = 19, monthly = 14, no-quiz= 23) and 70 on

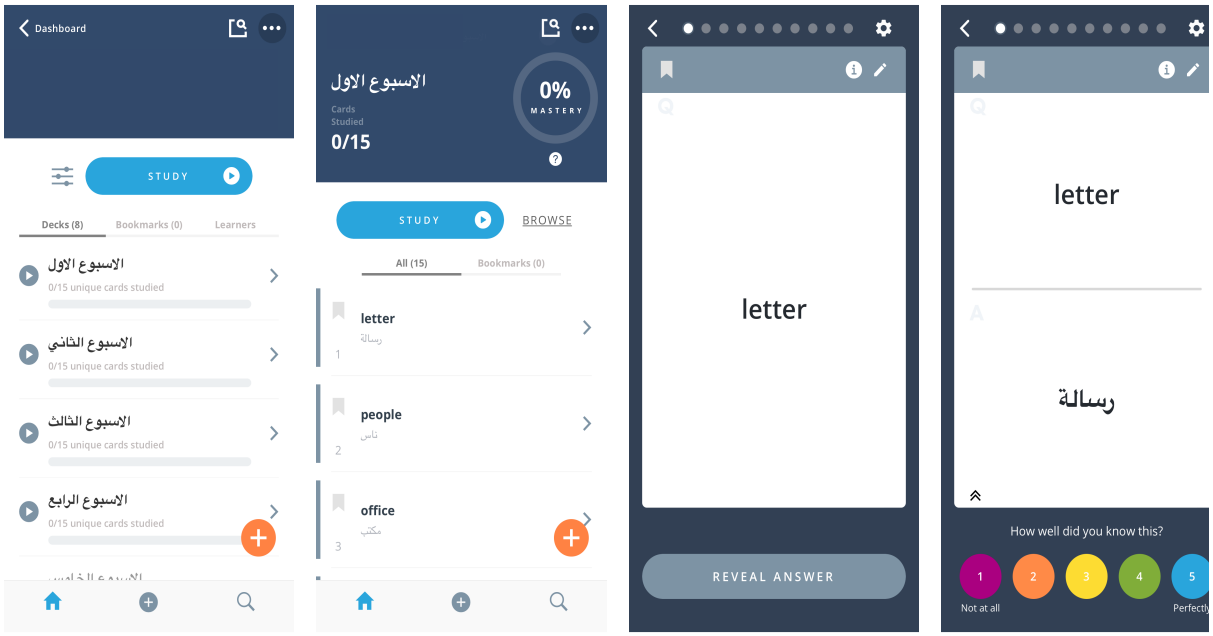
meaning recognition (weekly = 18, biweekly = 15, monthly = 14, no-quiz = 23). The dataset with the students who took both the pretests and posttests (i.e., complete cases) was used to answer the first two research questions given that time (pretest/posttest) is important to test the effectiveness of the intervention while the larger dataset was used to answer the remaining research questions to retain more statistical power.

The students receive five fifty-minute English classes per week and they have been learning English for seven years. All classes were identical in that they were taught by the same teacher, had the same textbooks and followed the same learning curriculum. An ethical approval from the University of Birmingham was granted before the study was conducted.

5.2.3 The digital flashcard platform

Conveniently, the majority of the effective techniques for learning from flashcards such as using retrieval and spaced repetitions (Nation, 2022) are often applied automatically in digital flashcards (Nakata, 2011). All four groups learned the target words through a digital flashcard platform called Brainscape using their phones (see Figure 12 for the app interface).

Figure 12. The interface of Brainscape



The choice of the app was based on app criteria and research criteria. In terms of the app criteria, the app needed to meet most of the criteria suggested by Nakata (2011). Nakata reviewed a number of flashcard software programs based on 17 criteria, 6 of which were related to flashcard creation and editing and 11 were related to learning. The criteria are shown in Table 15 below. As can be seen from the table, Brainscape checked most of the boxes indicating that it is an effective flashcard learning tool. In terms of the research criteria, Brainscape has an advantage over other commonly used apps such as Anki and Quizlet in that virtual classrooms can be created where researchers can remotely add, edit or remove items. The app also provides key information regarding students learning (e.g., how much time did students spend learning in minutes, the number of days spent learning, the number of unique words studied, the total number of words studied and mastery percentage level). A further advantage is that the app is free, unlike Anki which has no free version for iPhone operating system users.

Four virtual classes were created on Brainscape for each treatment group where their activities were monitored (mainly tracking the time spent learning). The flashcards were prepared by the researcher. Students were given instructions on how to access and use the Brainscape app. Additionally, they were given a WhatsApp number for contact in case they faced any technical difficulties with learning from the app.

Table 15. Criteria for evaluating flashcard software used to evaluate Brainscape (Nakata, 2011)

Criterion	Description	
	Flashcard creation and editing	
Flashcard creation	Can learners create their own flashcards	✓
Multilingual support	Can the target words and their translations be created in any language?	✓
Multi-word units	Can flashcards be created for multi-word units as well as single words?	✓
Types of information	Can various kinds of information be added to flashcards besides the word meanings (e.g., parts of speech, contexts, or audios)?	✓
Support for data entry	Does the software support data entry by automatically supplying information about lexical items such as meaning, parts of speech, contexts, or frequency information from an internal database or external resources?	?
Flashcard set	Does the software allow learners to create their own sets of flashcards	✓
	Learning	
Presentation mode	Does the software have a presentation mode, where new items are introduced and learners familiarize themselves with them?	✓
Retrieval mode:	Does the software have a retrieval mode, which asks learners to recall or choose the L2 word form or its meaning?	✓
Receptive recall	Does the software ask learners to produce the meanings of target words?	✓
Receptive recognition	Does the software ask learners to choose the meanings of target words?	x
Productive recall	Does the software ask learners to produce the target word forms corresponding to the meanings provided?	✓
Productive recognition	Does the software ask learners to choose the target word forms corresponding to the meanings provided?	x
Increasing retrieval effort	For a given item, does the software arrange exercises in the order of increasing difficult	✓

Generative use	Does the software encourage generative use of words, where learners encounter or use previously met words in novel contexts?	x
Block size	Can the number of words studied in one learning session be controlled and altered?	?
Adaptive sequencing	Does the software change the sequencing of items based on learners' previous performance on individual items?	✓
Expanded rehearsal	Does the software help implement expanded rehearsal, where the intervals between study trials are gradually increased as learning proceeds?	✓

Note. ✓ = feature exists, x = absent,? = partially exists in Brainscape.

5.2.4 Target words and tests

The students learned 120 words, which seems a reasonable goal to learn during a school semester (within a three-semester system school year) given the 100 word semester gains of same-level students from the first study and the reported 400 words average annual gains of EFL learners (Webb & Chang, 2012).

Although results from study 1 suggest that Saudi EFL students of a comparable level were not able to master the highest frequency words (1000 words), they are likely to know some of these words. If students by chance happen to know a large proportion of the target words then there might not be enough room for learning to occur. Therefore, this study included both high frequency words (1000-3000 frequency bands) and middle frequency words (4000 frequency band) to avoid a scenario where students score high on the pretests leaving little room for learning on the posttests.

One approach to choosing target words from a frequency band is to choose them randomly and then construct a self-made test to measure the learning of these words. However, this might

raise concerns regarding the validity and reliability of the self-made tests. Another more reliable and efficient approach is to use words from a validated test such as The UVLT (Webb et al., 2017). This has two advantages. First, the words in such tests are usually carefully selected to represent as much as possible the words in a frequency band (e.g., matching the percentage of nouns and verbs in the test with their percentage in the frequency band). Second and more importantly, an already validated test exists for these words which is likely to be more reliable than a self-made test.

Therefore, the 120 target words used in the present study come from the 1000-4000 levels of the UVLT (version B) which has been validated and used in several studies including study 1 (Chapter 4). The UVLT and a recall version created from the test were used as pre and posttests in this study to measure students' meaning recall and meaning recognition knowledge of the target words (Appendix 1 and Appendix 2). In other words, the UVLT served both as the test and its items as the target words in this study. Previous research has shown that bilingual vocabulary tests provide more reliability than monolingual ones, especially with low proficiency learners (Elgort, 2013). Thus, an Arabic version of the UVLT, which has been translated and reviewed by two translators with graduate degrees in translation, was used. The same study also found that cognates and loanwords can influence test scores significantly, thus they were replaced with non-cognate words from version A of the UVLT. Without this, it would not be possible to distinguish correct scores due to the intervention from correct scores due to the facilitative learning effect of loanwords (see section 2.2.1). The replacement was a cluster for a cluster (three target words with their six options, similar to Figure 4) instead of words for words to preserve the test structure and minimize changes (potential cognates identified: photograph, center, check, coach, weed, cap, super, junior, regime, vitamin, cave, scenario, soap, orchestra and tobacco). Cognates were not removed in study 1 because the aim

of that study was to measure learners vocabulary knowledge which loanwords are part of, hence they should be represented in the vocabulary knowledge estimates (Nation & Webb, 2011).

5.2.5 Self-regulation, motivation and learners' perceptions

The same instruments used to measure self-regulation and motivation in the first study were used here. The SRCvoc instrument (Tseng et al., 2006) was used to measure students' self-regulation capacity in learning vocabulary (Appendix 3). The instrument uses 6-point Likert-scale responses ranging from "strongly agree" to "strongly disagree". A score of 4 ("slightly agree") or above on any item indicates the possibility that a student has control over that dimension. The overall score of self-regulating capacity is obtained by calculating the total score of individual items. To measure motivation, the study used the STD-L2 (Alamer, 2021a), given that it has been validated in the Saudi context and found to be a reliable measure of L2 motivation (Alamer, 2021b). The SDT-L2 (Appendix 4) has 20 items with a 5-point Likert-type response format ranging from "strongly agree" to "strongly disagree". Unlike self-regulation, the motivation questionnaire does not have one overall score but consists of two main scales (autonomous and controlled motivation) which were used in the analysis.

A questionnaire measuring students' perceptions of frequent quizzing and digital flashcard learning was created (Appendix 5). The questionnaire includes three main scales: perceived effectiveness (students' beliefs about the effectiveness of frequent quizzes and flashcard learning), enjoyment (students' feelings about the quizzes and app), and future app use (students' intention to continue using the app). Perceived effectiveness and enjoyment scales each have six items (three for frequent quizzes and three for flashcard learning from the app) while future app use has three items. The instrument uses 5-point Likert-scale responses ranging from "strongly agree" to "strongly disagree". All instruments were translated into Arabic and reviewed by two translators with graduate degrees in translation.

5.2.6 Treatment

The four classes were assigned randomly to one of the four treatment groups shown in Figure 13. Students either took quizzes weekly, biweekly, monthly or no quizzes. The quizzes were in the form of receptive meaning recognition (multiple choice) where the English words were supplied and students chose the equivalent Arabic word. Meaning recognition tests were chosen since they are more objective than meaning recall questions, take less class time to complete and are quicker to grade. The weekly group quizzes involved 15 multiple-choice questions testing all the 15 words introduced every week. Similarly, the quizzes of the biweekly and monthly groups included 15 questions but covered a random selection of all the untested words introduced to that point. One criterion for word selection for learning is frequency. For example, of the 15 weekly words to be learned, 3-4 were selected from each of the four frequency levels (1000-4000) so that every week students learn higher and lower frequency words equally.

Figure 13. Types of treatments

Class	Quizzes frequency	Number of total quizzes	Number of words on each quiz	Number of words covered in each quiz	Credit per quiz
1	Weekly	8	15	15	2%
2	Biweekly	4	15	30	4%
3	Monthly	2	15	60	8%
4	No-quiz	0	0	0	0

All the groups were required to do grade-related tasks totaling 16 course grades (i.e., 16% of total class credit). In the case of the weekly group, every individual quiz was worth two grades, four for the biweekly group and eight for the monthly group. Although the last group was not required to take quizzes, they were required to access the digital flashcard app once a week to

earn two grades. This is because a previous study found that two thirds of a class did not access a digital flashcards app even once when it was completely optional (Seibert Hanson & Brown, 2019).

5.2.7 Procedure

Figure 14 presents an overview of the study procedure. In the first week of the semester, all students took the UVLT in both meaning recognition and meaning recall formats to control for their prior knowledge of the 120 target words. In addition, they completed the motivation and self-regulation questionnaires. In the same week, all groups were informed about the importance of high frequency vocabulary and the efficiency of learning from flashcards to justify the effort they will put in throughout the semester as well as instructions on how to access the digital flashcards. In the second week, actual learning began lasting for eight school weeks. A three-week break occurred between the third and fourth weeks of the experiment. Students were informed of the quizzes and their frequency beforehand. They were given 15 minutes to complete them. Once they were finished, the students were asked to hand in the quizzes to the teacher and no feedback was given. On the tenth week, unannounced posttests which were identical to the pretests were administered along with the perceptions questionnaire. All tests and questionnaires were given in class in paper format.

Figure 14. Procedure overview

School Weeks	Weekly group	Biweekly group	Monthly group	No-quiz group
1	Vocabulary pretests Motivation and self-regulation questionnaires			
2	Quiz 1			
3	Quiz 2	Quiz 1		
Break week 1				
Break week 2				
Break week 3				
4	Quiz 3			
5	Quiz 4	Quiz 2	Quiz 1	
6	Quiz 5			
7	Quiz 6	Quiz 3		
8	Quiz 7		Quiz 2	
9	Quiz 8	Quiz 4		
10	Perceptions questionnaire Unannounced vocabulary posttests			

5.2.8 Analysis

The results of the four groups on the frequent quizzes, vocabulary tests and questionnaires were analyzed using GLMM (Baayen et al., 2008) to answer the five research questions. GLMM can incorporate both fixed and random effects in one model which is particularly useful when individual variation between subjects is expected (Linck, 2016).

The dependent variables in this study were the UVLT and recall UVLT which were scored dichotomously (0 incorrect and 1 correct). For recall UVLT, 1 point was given if exact or close meaning was provided. For a small number of cases (mainly unclear handwriting), another rater was consulted, and an inter-rater agreement was reached on all of these cases. The models in this study were fitted with the subject and item (UVLT words) as random effects while time (pretest vs. posttest), group (weekly, biweekly, monthly and no-quiz), motivation and self-regulation were fixed effects. To ensure an inclusive analysis of the individual differences, separate models were constructed for motivation and self-regulation before combining them in a comprehensive model. The analysis was conducted on the overall scores of self-regulation. Motivation does not have a single overall score but comprises two main scales (autonomous and controlled motivation) which were used in the analysis. For perceptions, each one of the five scales (quiz joy, quiz effectiveness, app joy, app effectiveness and future use) was analyzed separately. The models reported were chosen following a forward selection approach based on likelihood ratio tests.

To test the differences between groups on the motivation, self-regulation and perceptions scales, a non-parametric ANOVA (i.e., The Kruskal–Wallis test) was computed given that the questionnaires did not meet the assumptions of one-way ANOVA.

5.3 Results

Table 16 shows the reliability of the two tests and the three questionnaires used in the study. All the instruments had good reliability with Cronbach alpha scores above 0.80.

Table 16. Reliability scores of the instruments used in the study.

Instrument	Items	Cronbach α
UVLT pretest	120	.98
UVLT posttest	120	.98
Recall UVLT pretest	120	.98
Recall UVLT posttest	120	.99
Autonomous motivation	10	.91
Controlled motivation	10	.82
Self-regulation	20	.80
Perceptions	15	.81

5.3.1 RQ1. Do frequent vocabulary quizzes (in any frequency) lead to globally more vocabulary learning compared to no quizzes?

The first research question asked whether the mere existence of quizzes leads to more vocabulary learning than no quizzes at all. To answer this question, the three quizzed groups (weekly, biweekly and monthly) were combined into one group (quiz group) and their results were compared to the no-quiz group.

The descriptive statistics in Table 17 summarize these results and show that the two groups had very similar results on the recognition pretest. The average of the no-quiz group was 23.74 while the quiz group average was 23.11. The scores of the two groups however differed on the posttest with the no-quiz showing virtually no difference (23.35) and appearing to make no gains (-0.39). In contrast, the quiz group scores increased on the posttest (37.51) and made relatively large gains (14.40) compared to the no-quiz group. In terms of the meaning recall test, the no-quiz group started with slightly lower scores on the pretest (5.52) compared to the

quiz group (8.06). The scores of both groups increased on the posttest with the quiz group scoring higher (14.08) than the no-quiz group (6.43). The gains of the quiz group were larger (6.02) than the no-quiz group (0.91).

Table 17. Meaning recognition and meaning recall tests score for the quiz and no-quiz groups

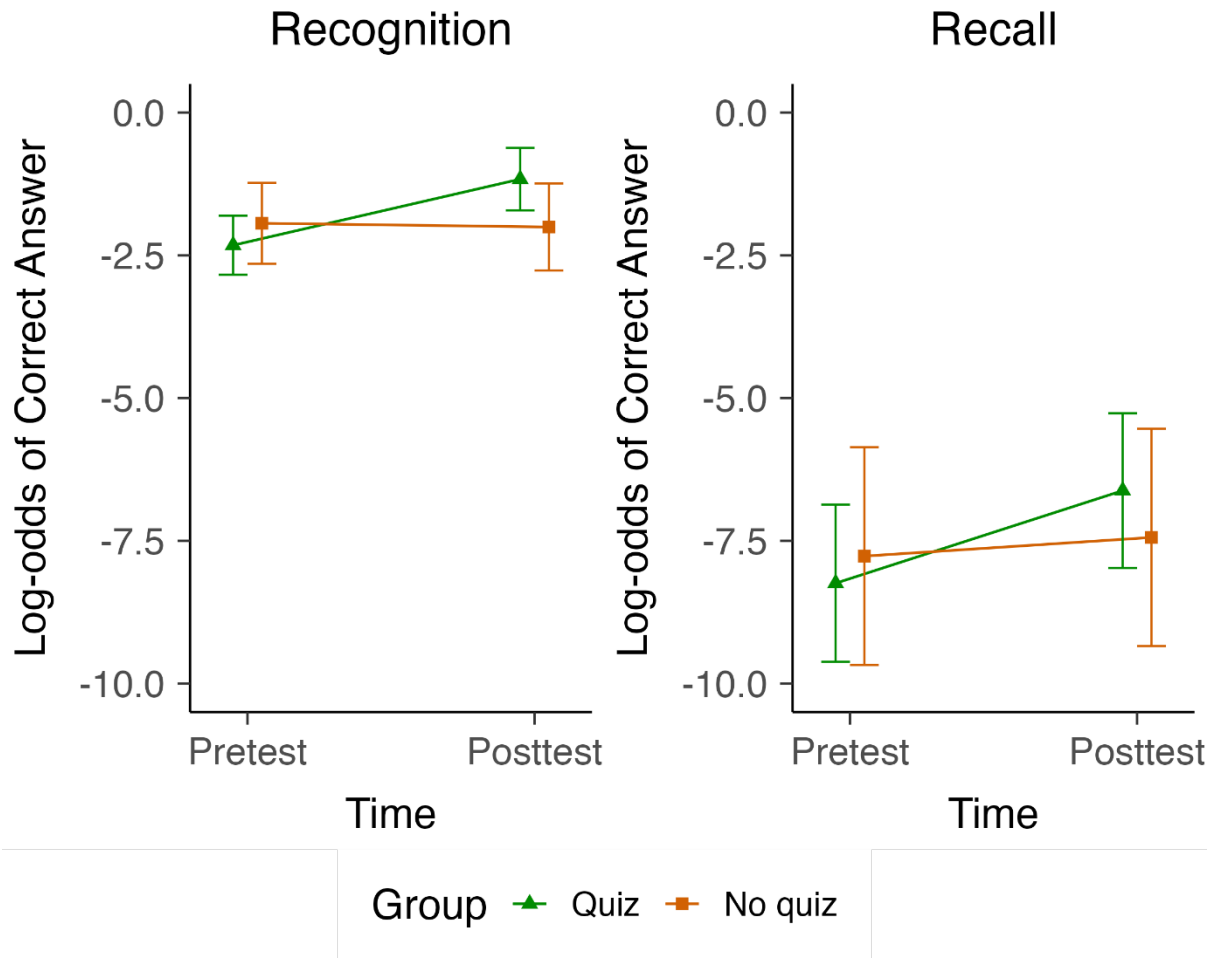
		Pretest			Posttest			Gain		
		M	SD	CI	M	SD	CI	M	SD	CI
Recognition	No-quiz	23.74	15.17	[23.17, 24.31]	23.35	18.41	[22.66, 24.03]	-0.39	9.18	[-4.36, 3.58]
	Quiz	23.11	26.78	[22.41, 23.81]	37.51	34.24	[36.62, 38.40]	14.40	18.12	[9.09, 19.72]
Recall	No-quiz	5.52	9.80	[5.16, 5.89]	6.43	10.30	[6.05, 6.82]	0.91	2.37	[-0.11, 1.94]
	Quiz	8.06	15.05	[7.69, 8.43]	14.08	26.14	[13.43, 14.72]	6.02	14.80	[1.94, 10.10]

Note. CI = 95% confidence interval

To test the significance of these observations, mixed logistic models were fitted with subject and item as random effects and time of test (pretest and posttest) and group (quiz and no-quiz, with the no-quiz being the baseline), and the interaction between the two as fixed effects. Adding by-subject random slopes for time improved model fit significantly. The comparability of the two groups was checked and the pretest results showed no significant difference between the quiz and the no-quiz group on meaning recognition ($b = -0.43$, $z = -0.91$, $p = .358$) and meaning recall ($b = 0.38$, $z = 0.32$, $p = .745$).

For meaning recognition knowledge, results showed no simple effects of test time ($b = -0.06$, $z = -0.27$, $p = .782$) but did show a significant interaction between test time and group ($b = 1.22$, $z = 4.18$, $p < .001$). Pairwise comparisons using the emmeans package (with Bonferroni adjustment for multiple comparisons) showed that the quiz group scored significantly higher on the posttest than the pretest ($b = 1.94$, $z = 4.60$, $p < .001$) but the no-quiz group did not ($b = 0.67$, $z = -1.76$, $p = .078$) (see Figure 15 left panel). In terms of meaning recall, a similar model to the meaning recognition was fitted. Results were also similar in that there were no simple effects of test time ($b= 0.32$, $z = 1.84$, $p = .064$) but a significant interaction between test time and group ($b = 1.29$, $z = 6.26$, $p < .001$). Pairwise comparisons showed that the quiz group scored significantly higher on the posttest ($b = -1.62$, $z = -14.84$, $p < .001$) but not the no-quiz group ($b = -0.32$, $z = -1.84$, $p = .387$) (see Figure 15 right panel). Having established that quizzes lead to significantly more vocabulary learning to the meaning recognition and meaning recall level compared to no quizzes, the second research question investigates the effect of different quiz frequencies on vocabulary learning.

Figure 15. Comparing the effect of quizzes on meaning recognition and meaning recall vocabulary learning from digital flashcards.



5.3.2 RQ2. Which of the three types of quizzes frequency (weekly, biweekly, monthly) result in more vocabulary learning by the end of the semester as measured by a posttest?

Table 18 and 19 summarize the results of meaning recognition and meaning recall tests respectively for the three quizzed groups (weekly, biweekly and monthly). For meaning recognition, the biweekly group average score on the pretest was slightly lower (18.07) than the weekly (24.39) and monthly groups (26.86). This variation however diminished when the gains of the three groups are examined. Results showed similar gains for all groups with an average gain of 14.93 for the weekly group, 13.93 for the biweekly group and 14.39 for the monthly group. For meaning recall, the biweekly group score on the pretest was lower (3.74) than the weekly (11.45) and monthly (9.07) groups. The scores of all groups increased on the posttest with the weekly group scoring the highest (16.10) followed by the biweekly group (13.37) which had similar posttest scores to the monthly group (12.14). The gains of the biweekly group (9.63) were larger than the gains made by the weekly (4.65) and monthly (3.07) groups.

Table 18. Meaning recognition tests score

	Pretest			Posttest			Gain		
	M	SD	CI	M	SD	CI	M	SD	CI
Weekly	24.39	27.42	[10.75, 38.02]	38.78	40.35	[18.71, 52.72]	14.39	16.30	[6.28, 22.50]
Biweekly	18.07	16.37	[9.00, 27.13]	32.00	33.09	[13.68, 50.32]	13.93	20.98	[2.31, 25.55]
Monthly	26.86	35.69	[6.25, 47.47]	41.79	29.53	[24.73, 58.84]	14.93	18.41	[4.30, 26.56]

Note. CI = 95% confidence interval

Table 19. Meaning recall tests score

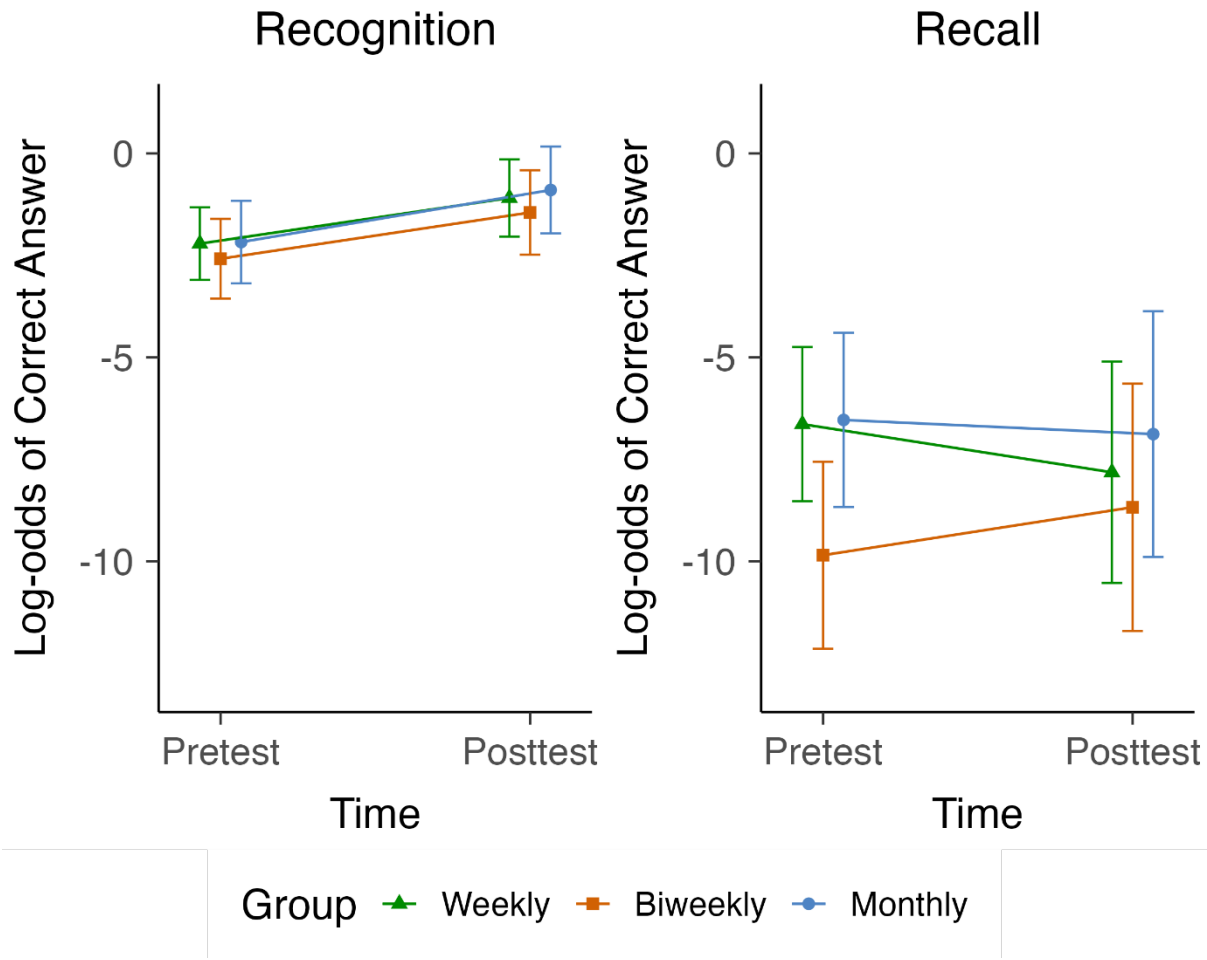
	Pretest			Posttest			Gain		
	M	SD	CI	M	SD	CI	M	SD	CI
Weekly	11.45	17.88	[3.08, 19.82]	16.10	28.90	[2.57, 29.36]	4.65	11.68	[-0.82, 10.12]
Biweekly	3.74	8.56	[-0.39, 7.86]	13.37	29.08	[-0.65, 27.38]	9.63	21.02	[-0.50, 19.76]
Monthly	9.07	17.55	[-1.06, 19.21]	12.14	19.55	[0.86, 23.43]	3.07	5.85	[-0.31, 6.45]

Note. CI = 95% confidence interval

Before testing the significance of these observations, the comparability of the groups was checked through a logistic model and the results showed no significant difference between the groups on the meaning recognition pretest (all ps >.05). In terms of meaning recall, the score of the biweekly group on the pretest was significantly lower than the weekly (baseline) group (b = -4.13, z = -2.87, p = .004), while no significant differences existed between the weekly and monthly groups.

For meaning recognition knowledge, results showed a significant simple effect for test time (b = 1.11, z = 3.47, p < .001) and no significant interaction between the posttest and the biweekly group (b = -0.37, z = -0.56, p = .575) or the posttest and the monthly group (b = -0.03, z = -0.05, p = .954). Pairwise comparisons showed significant vocabulary learning for all groups with the log-odds of the weekly (b = -1.47, z = -4.45, p = .007), biweekly (b = -1.13, z = -3.18, p = .021) and monthly groups (b = -1.27, z = -3.48, p = .007) increasing significantly on the posttest. These findings suggest that the gains made by groups with fewer quizzes, namely the biweekly and monthly groups, were not significantly different from the gains made by the weekly group (see Figure 16, left panel). For meaning recall knowledge, results showed no significant simple effect of test time (b = -1.18, z = -1.79, p = .072) or significant interaction between the posttest and the monthly group (b = 1.36, z = 1.61, p = .107). As mentioned earlier, the biweekly group showed a significant difference on the pretest (b = -3.21, z = -2.36, p = .020) from the weekly group and also showed a significant interaction with test time (b = 2.35, z = 2.86, p = .004) indicating that their scores were lower on both the pretest and posttest than the weekly group. Pairwise comparisons indicated that none of the three groups showed a significant difference in their scores on the posttest (all ps >.05). These findings of meaning recall tests show that although the biweekly group scores were significantly lower on both the pretest and posttest from the weekly group, all groups were similar in that none of them improved significantly from the pretest to the posttest (see Figure 16, right panel).

Figure 16. Comparing the effects of the three types of quiz frequency on meaning recognition and meaning recall vocabulary learning.



An additional analysis was conducted on the meaning recognition and meaning recall vocabulary gains using linear models. The gain of every student was calculated by subtracting the posttest score from the pretest score (i.e., gain = posttest – pretest). The findings were similar to the results of the mixed effects models reported above (see Appendix 6). The no-quiz group scores were significantly lower than the weekly (reference) group ($b = -14.78$, $z = -2.93$, $p = .005$). Meanwhile, the gains of the biweekly and monthly groups were not significantly different from the weekly group (all $p > .05$). For meaning recall, no significant differences were found across all groups (all $p > .05$).

Taken together, the findings suggest no advantage for groups with more frequent vocabulary quizzes in meaning recognition and meaning recall vocabulary learning. All groups made significant vocabulary gains on meaning recognition knowledge, and, at the same time, all groups showed no significant learning to the meaning recall level.

5.3.3 RQ.3 What role do individual differences, namely motivation and self-regulation, play in out-of-class vocabulary learning from digital flashcards?

The section starts by analyzing students' responses to the motivation and self-regulation questionnaires and explores if the groups differ significantly from one another in their responses. This is followed by logistic modeling to examine the effects of motivation and self-regulation on meaning recognition and meaning recall vocabulary learning.

5.3.3.1 Questionnaires analysis

Table 20 shows the level of autonomous motivation, controlled motivation and self-regulation of students in each group. The biweekly group had higher average scores on all measures than the other groups. They demonstrated the highest autonomous motivation (4.28), controlled motivation (4.87) and self-regulation (4.00) levels.

To test the significance of these differences and other differences, the Kruskal–Wallis test was computed. For autonomous motivation, results showed a significant main effect for group (H (3) = 6.20, p <.001). Pairwise comparison (Dunn's Test) with Bonferroni adjustment for multiple comparisons showed that the biweekly group had significantly higher autonomous motivation level than the weekly (H = 69.54, p <.001) and monthly (H = 44.13, p <.001) groups but not the no-quiz group (H = 24.72, p =.555). The no-quiz group had significantly higher levels of autonomous motivation than the weekly group (H = 44.81, p =.014) but not the monthly group (H = 19.41, p = 1.000).

Table 20. The levels of autonomous motivation, controlled motivation and self-regulation for each group.

	Autonomous motivation		Controlled motivation		Self-regulation	
	M	SD	M	SD	M	SD
Weekly	3.83	0.65	3.41	0.66	4.53	0.91
Biweekly	4.28	0.53	4.00	0.66	4.87	0.69
Monthly	3.80	1.05	3.73	0.73	4.56	1.02
No-quiz	4.04	0.84	3.58	0.92	4.73	0.57

Note. Max score is 5 for autonomous motivation and controlled motivation. Max score is 6 for self-regulation.

For controlled motivation, results showed a significant main effect for group (H (3) = 35.42, p <.001). Pairwise comparison showed that the biweekly group had significantly higher controlled motivation level than the weekly (H = 84.73, p <.001) and the no-quiz (H = 64.92, p <.001) groups but not the monthly group (H = 37.44, p =.099). The monthly group had significantly higher levels of controlled motivation than the weekly group (H = 47.28, p =.015) but not the no-quiz group (H = 19.41, p = 1.000).

Unlike the previous two scales, group scores on the self-regulation questionnaire were quite similar with the biweekly (4.87) and no-quiz (4.73) groups being more similar and higher than the weekly (4.53) and monthly groups (4.56). The results of a Kruskal–Wallis test showed no significant main effect for group ($H(3) = 4.76$, $p = .123$) on self-regulation levels.

5.3.3.2 Separate mixed effects models

Similar to the first study, separate models were constructed for motivation and self-regulation before combining them into a comprehensive model. All of the models had the same base model which consisted of fixed effects of time and group and an interaction between the two, subject and item as random effects and by-subject random slopes for time. The dependent variable in all models was the pretest and posttest scores. In the separate analysis, all predictors were fitted in models first without interactions, next in two-way interactions with group and finally in three-way interactions with group and time (three-way interactions that did not improve model fit significantly are not reported). The separate and comprehensive analyses were conducted twice, once for meaning recognition knowledge and once for meaning recall knowledge.

Starting with meaning recognition, a model was fitted to examine the effect of autonomous motivation on student test scores. The autonomous motivation model had a significantly better fit compared to the base model ($\chi2\ (1) = 12.12$, $p < .001$). The model fit improved further with a two-way interaction between autonomous motivation and group ($\chi2\ (1) = 8.50$, $p = .036$). The simple effect of autonomous motivation on the meaning recognition test scores was not significant ($b = -0.54$, $z = -0.54$, $p = .195$), however the interactions between autonomous motivation and the biweekly ($b = 1.82$, $z = 2.95$, $p = .003$), monthly ($b = 1.49$, $z = 2.90$, $p = .003$) and no-quiz ($b = 1.35$, $z = 2.52$, $p = .011$) groups were significant. Simple slope analysis (using the interactions package) and visual inspection of the interaction (see Figure 17, left

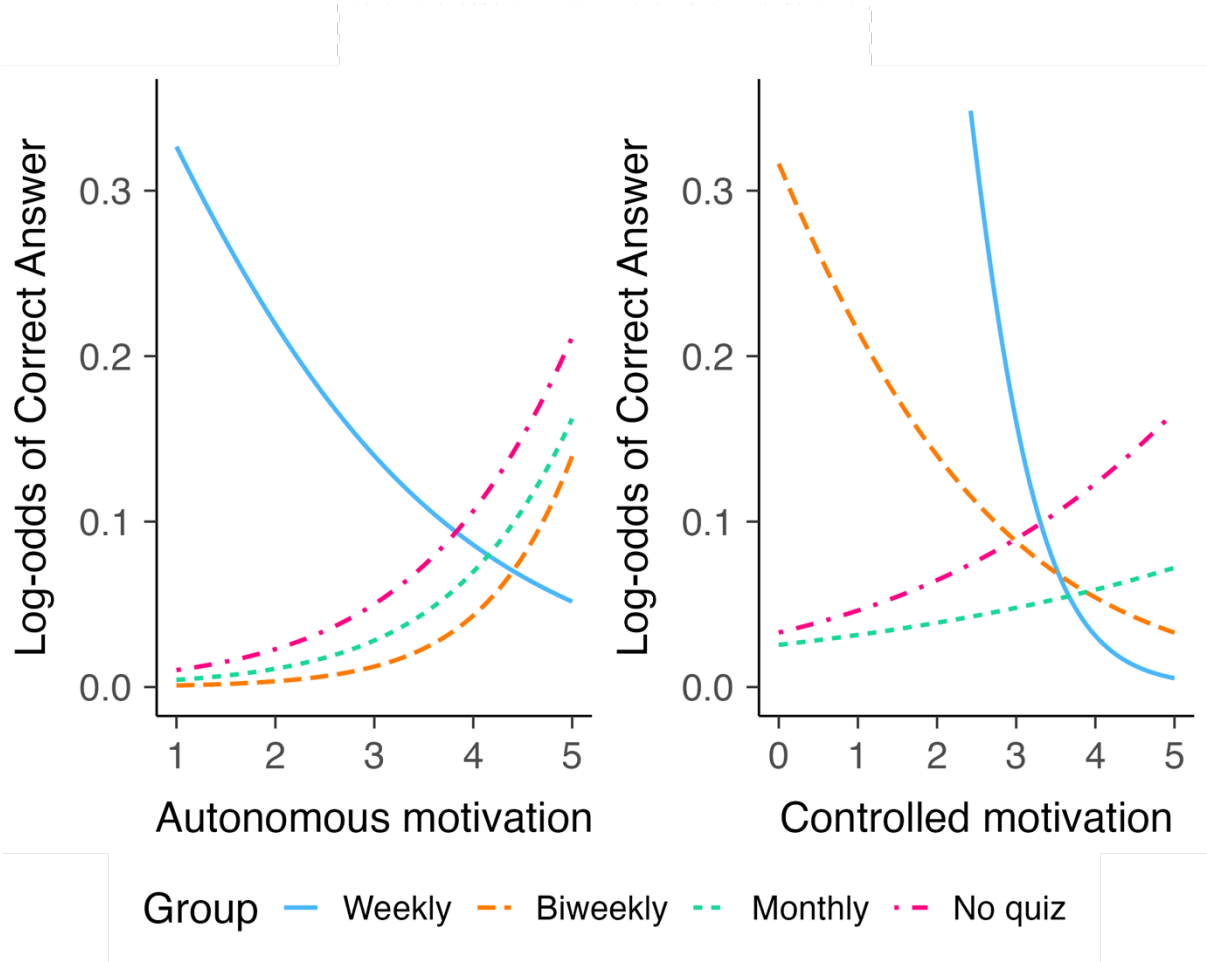
panel) showed that autonomous motivation had a significant positive effect on meaning recognition test scores for the biweekly (b = 1.27, z = 2.47, p = .013), the monthly group (b = 0.95, z = 3.05, p = .002) and the no-quiz group (b = 0.81, z = 2.30, p = .021) but not the weekly group (b = -0.54, z = -1.15, p = .247).

Controlled motivation, the second half of motivation in this study, did not improve model fit compared to the base model ($\chi 2$ (1) = 0.84, p = .358). However, the model fit improved significantly with a two-way interaction between controlled motivation and group ($\chi 2$ (3) = 13.98, p = .002). Results showed a negative simple effect of controlled motivation on the meaning recognition test scores (b = -1.78, z = -3.72, p < .001) and significant interactions between controlled motivation and the biweekly (b = 1.82, z = 2.95, p = .003) and monthly groups (b = 1.49, z = 2.90, p = .003) but not the biweekly group (b = 1.26, z = 1.92, p = .054). Simple slope analysis and visual inspection of the interaction (see Figure 17, right panel) showed that controlled motivation had a significant negative effect on meaning recognition test scores for the weekly group only (b = -1.78, z = -3.65, p < .001) while no significant effect was found for the other three groups (All ps > .05).

Self-regulation did not have a significant effect on students' meaning recognition test scores ($\chi 2$ (1) = 1.17, p = .278), or its interaction with group ($\chi 2$ (1) = 1.36, p = .713).

Overall, the analysis of recognition vocabulary tests shows that students with higher autonomous motivation in the biweekly, monthly and no-quiz groups scored higher on the recognition tests but not the weekly group. In contrast, higher levels of controlled motivation had a significant negative effect on the test scores of the weekly group only. Finally, self-regulation did not have a significant effect on test scores.

Figure 17. Interactions plots for the effect of autonomous motivation (left) and controlled motivation on meaning recognition (right).

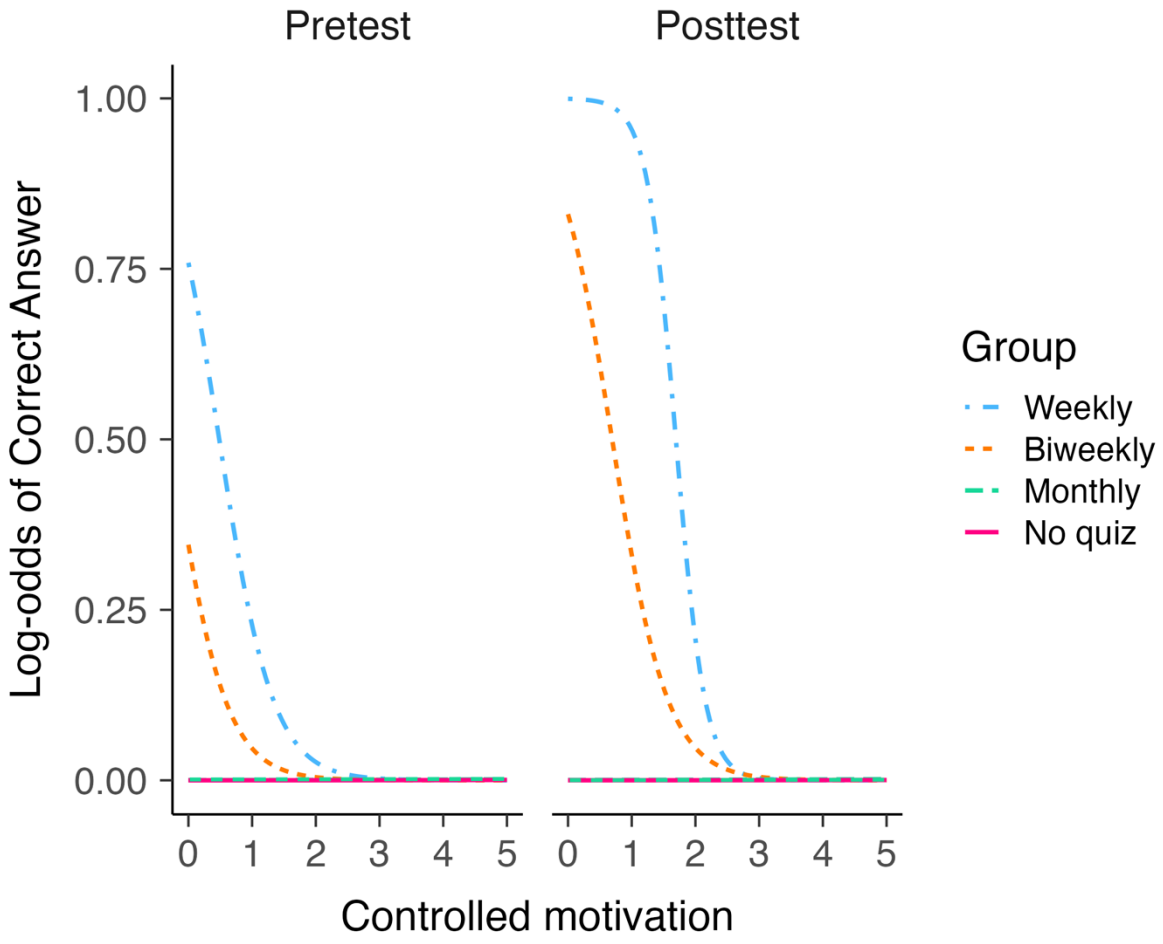


In terms of meaning recall vocabulary knowledge, a model was fitted to investigate the effect of autonomous motivation on meaning recall test scores. The model had a significantly better fit compared to the base model ($\chi 2$ (1) = 23.84, $p < .001$), but did not improve significantly with a two-way interaction ($\chi 2$ (3) = 1.38, $p = .710$). The results of the model showed that higher levels of autonomous motivation significantly increased the log-odds of correct answers on the tests for all students regardless of group ($b = 2.31$, $z = 4.49$, $p < .001$).

Controlled motivation, on the other hand, did not have a significant effect on students' meaning recall test scores with or without two-way interaction (all ps >.05). However, a model with a three-way interaction with group and time improved model fit significantly ($\chi 2$ (8) = 17.10, p

= .029). Results showed significant three-way interactions between controlled motivation, time and the weekly (b = -2.01, z = -3.63, p < .001), biweekly (b = 2.09, z = 2.31, p = .020) and the no-quiz groups (b = 1.89, z = 2.60, p = .009). Simple slope analysis of the three-way interactions showed that controlled motivation had a significant negative effect on the weekly group pretest (b = -2.37, z = -2.12, p = .034) and posttest (b = -4.38, z = -3.18, p = .001) (see Figure 18). Controlled motivation also negatively affected the test scores of the biweekly group on the pretest (b = -2.37, z = -1.99, p = .046) but not the posttest (b = -2.29, z = -1.61, p = .106). Finally, no significant effect was found for controlled motivation on the test scores of the monthly and no-quiz groups (All ps > .05)

Figure 18. Three-way interaction between controlled motivation, group and time on meaning recall test.



The model for the effect of self-regulation on students' meaning recall test scores had a significantly better fit compared to the base model ($\chi2$ (1) = 4.50, p = .033). Adding an interaction between self-regulation and group did not improve model fit significantly ($\chi2$ (3) = 2.23, p = .525). The results showed a positive significant effect for self-regulation on meaning recall test scores regardless of group (b = 1.06, z = 2.21, p = .026).

Overall, the analysis of meaning recall vocabulary test scores shows that students with higher autonomous motivation scored higher on the meaning recognition tests regardless of group. Controlled motivation had a significant three-way interaction with test time and group. Results showed that higher levels of controlled motivation had a significant negative effect on both the pretest and posttest scores of the weekly group. Results also showed that students in the biweekly group with higher controlled motivation scored significantly lower on the pretest scores but not the posttest. Finally, higher levels of self-regulation had a significant positive effect on test scores regardless of group.

5.3.3.3 Comprehensive models

All factors that were shown to be significant in the previous analysis were combined in comprehensive models to examine how they jointly affect meaning recognition and meaning recall vocabulary knowledge. The comprehensive models were then compared to the base models which consisted of the pretest and posttest scores as a dependent variable, time and group and an interaction between the two as fixed effects, subject and item as random effects and by-subject random slopes for time.

For meaning recognition vocabulary knowledge, the comprehensive model included an interaction between autonomous motivation and group and another interaction between controlled motivation and group. The model fit was a significant improvement compared to the base model ($\chi2$ (5) = 17.50, p = .003). The findings (Table 22) showed a negative simple effect

for controlled motivation on meaning recognition test scores (b = -2.00, z = -3.65, p < .001) and significant interactions between controlled motivation and the monthly (b = 1.80, z = 2.52, p = .011) and the no-quiz groups (b = 1.82, z = 2.76, p = .005). Simple slope analysis showed that controlled motivation had a significant negative effect on the test scores of the weekly (b = -2.10, z = -4.00, p < .001) and biweekly groups (b = -1.18, z = -2.67, p = .007) while no significant effects were found on the scores of the monthly (b = -0.18, z = -0.34, p = .661) and no-quiz groups (b = -0.18, z = -0.51, p = .607; see Figure 19). Autonomous motivation did not show a significant effect on students' meaning recognition test scores (b = 0.49, z = 0.93, p = .350) nor its interaction with group (all ps >.05). Overall, the comprehensive meaning recognition model showed that students in the weekly and biweekly groups with higher controlled motivation performed lower on the vocabulary test scores.

Figure 19. Interaction plot between controlled motivation and group in the comprehensive meaning recognition model

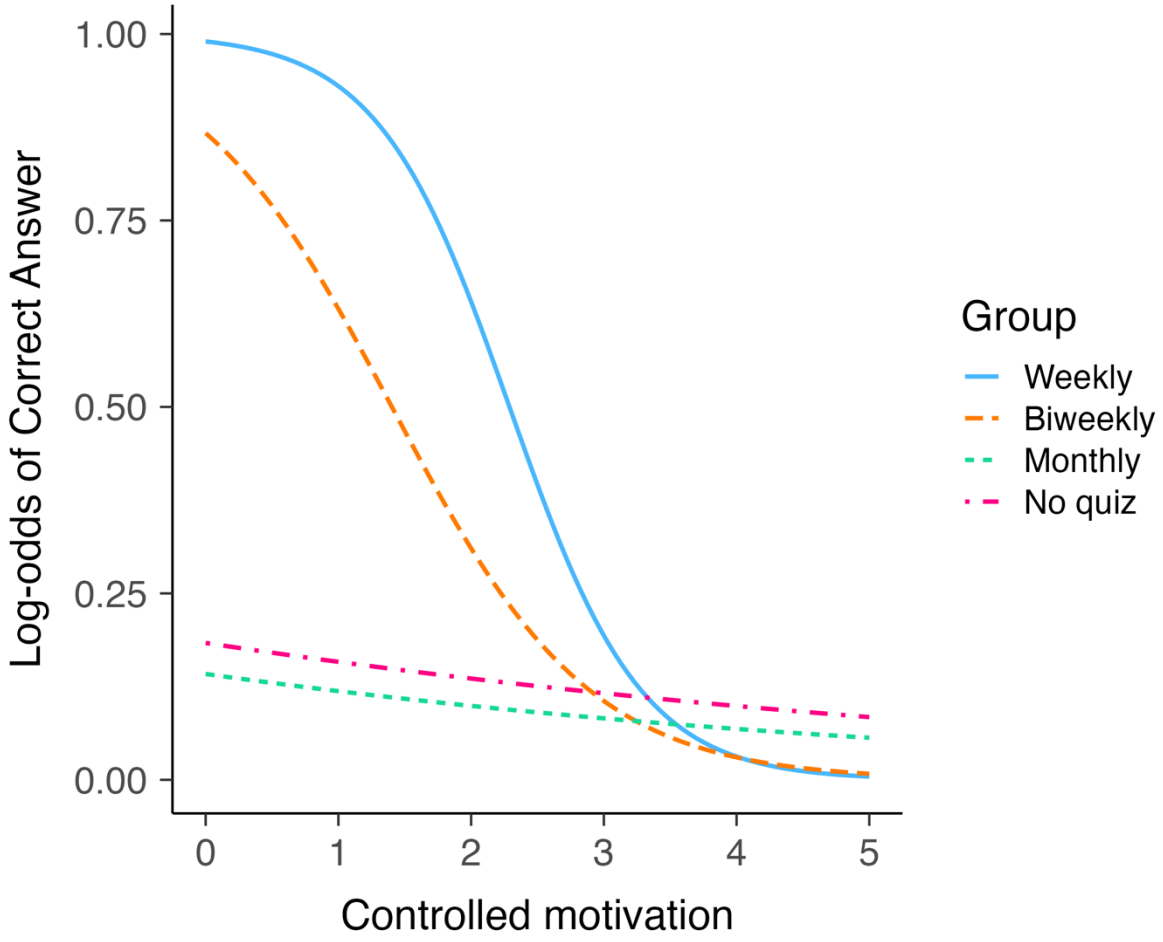


Table 21. Comprehensive mixed effects model output for meaning recognition vocabulary.

Fixed effects	β	Std. Error	Z value	P
Intercept	2.60	1.69	1.53	0.126
Posttest	1.24	0.28	4.40	**< 0.001**
Biweekly	-8.44	2.98	-2.84	**0.005**
Monthly	-8.37	2.37	-3.53	**< 0.001**
No-quiz	-7.81	2.14	-3.65	**< 0.001**
Autonomous motivation	0.50	0.53	0.93	0.350
Controlled motivation	-2.01	0.55	-3.66	**< 0.001**
Biweekly * Posttest	-0.03	0.41	-0.08	0.932
Monthly * Posttest	0.25	0.43	0.60	0.549
No-quiz * Posttest	-1.27	0.37	-3.44	**0.001**
Biweekly * Autonomous motivation	1.42	0.78	1.82	0.069
Monthly * Autonomous motivation	0.49	0.63	0.78	0.436
No-quiz * Autonomous motivation	0.43	0.67	0.64	0.521
Biweekly * Controlled motivation	0.67	0.73	0.92	0.355
Monthly * Controlled motivation	1.80	0.72	2.52	**0.012**
No-quiz * Controlled motivation	1.83	0.66	2.77	**0.006**

Random effects variance (subject = 2.14, item = 1.00)

For meaning recall vocabulary knowledge, the comprehensive model consisted of autonomous motivation, self-regulation and a three-way interaction between controlled motivation, time and group. The comprehensive model had a better fit compared to the base model ($\chi 2$ (7) = 15.20, p = .033). The results of the comprehensive model (Table 22) showed a significant main effect for autonomous motivation on meaning recall vocabulary knowledge. The log-odds of

correct answers on the test increased for students with higher autonomous motivation regardless of group ($b = 2.56$, $z = 4.43$, $p < .001$). Controlled motivation showed a negative and significant simple effect on meaning recall test scores ($b = -3.74$, $z = -3.62$, $p < .001$). The two-way interactions between controlled motivation and group were significant for the monthly ($b = 2.98$, $z = 1.97$, $p = .047$) and the no-quiz groups ($b = 2.64$, $z = 2.06$, $p = .039$) but not the biweekly group ($b = -0.10$, $z = -0.06$, $p = .950$). The three-way interactions were significant for all four groups suggesting that the effect of controlled motivation varies depending on test time (see Figure 20). Simple slope analysis of the three-way interactions showed that controlled motivation had a significant negative effect on the weekly group pretest ($b = -3.74$, $z = -3.62$, $p < .001$) and posttest ($b = -5.69$, $z = -4.22$, $p < .001$). Similarly, controlled motivation negatively affected both the pretest ($b = -3.84$, $z = -2.97$, $p = .003$) and posttest scores ($b = -3.49$, $z = -2.23$, $p = .025$) of the biweekly group. Finally, no significant effect was found for controlled motivation on the test scores of the monthly and no-quiz groups (All $ps > .05$). Overall, the comprehensive meaning recall model showed that students with higher levels of autonomous motivation scored higher on the vocabulary tests regardless of group. Also, it showed that students in the weekly and biweekly groups with higher controlled motivation performed lower on meaning recall test scores.

Figure 20. Three-way interaction between controlled motivation, group and time in the comprehensive meaning recall model.

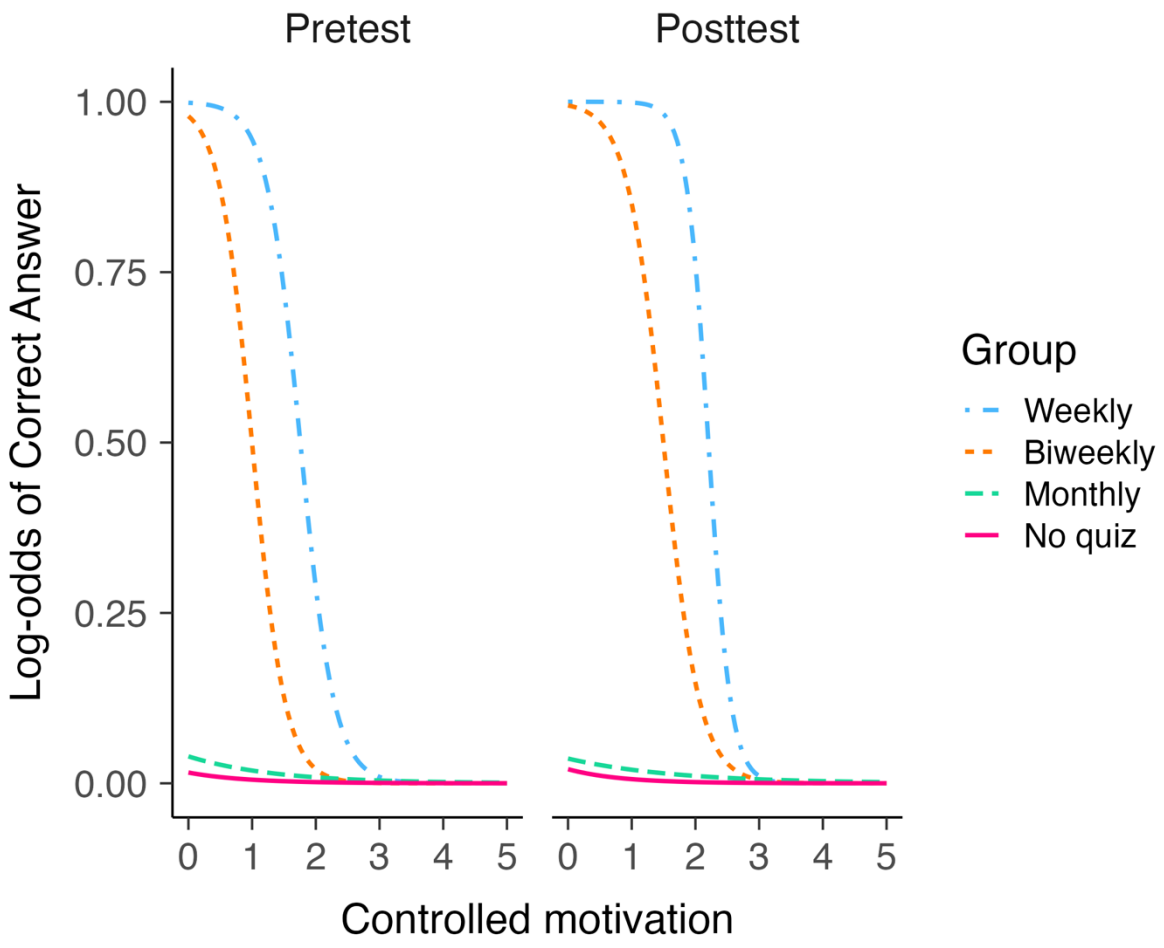


Table 22. Comprehensive mixed effects model output for meaning recall vocabulary.

Fixed effects	β	Std. Error	Z value	P
Intercept	-5.07	4.02	-1.26	0.208
Posttest	5.97	1.83	3.26	**0.001**
Biweekly	-2.74	5.98	-0.46	0.647
Monthly	-9.78	5.44	-1.80	0.072
No-quiz	-10.72	4.53	-2.37	**0.018**
Autonomous motivation	2.57	0.58	4.43	**< 0.001**
Self-regulation	0.28	0.63	0.44	0.656
Controlled motivation	-3.75	1.03	-3.62	**< 0.001**
Biweekly * Posttest	-4.58	3.53	-1.30	0.195
Monthly * Posttest	-6.06	3.81	-1.59	0.112
No-quiz * Posttest	-5.69	2.51	-2.27	**0.023**
Posttest * Controlled motivation	-1.95	0.57	-3.44	**< 0.001**
Biweekly * Controlled motivation	-0.10	1.62	-0.06	0.951
Monthly * Controlled motivation	2.98	1.51	1.98	**0.048**
No-quiz * Controlled motivation	2.65	1.28	2.06	**0.039**
Biweekly * Controlled motivation * posttest	2.30	1.00	2.30	**0.021**
Monthly * Controlled motivation * posttest	2.10	1.06	1.98	**0.048**
No-quiz * Controlled motivation * posttest	1.82	0.73	2.48	**0.013**

Random effects variance (subject = 8.75, item = 7.11)

Overall, results showed a significant and positive main effect for autonomous motivation on meaning recall but not meaning recognition. Moreover, controlled motivation had a significant negative effect only on the weekly and biweekly meaning recognition groups on both meaning

recognition and meaning recall tests. Finally, no significant effect was found for self-regulation on either meaning recognition or meaning recall.

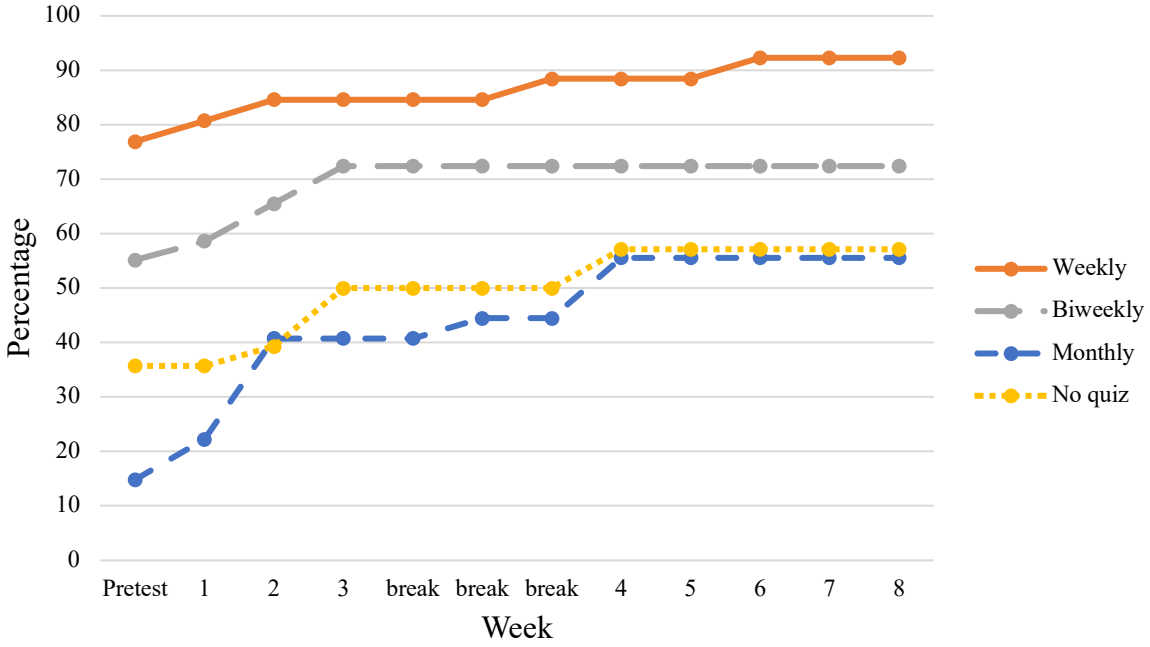
5.3.4 Students' activity on Brainscape and vocabulary learning

The previous sections focused on the direct effects of testing. In this section, the indirect effects of testing on vocabulary learning are discussed, that is, the effects of quizzes and quiz frequency on the learning behavior of students from the digital flashcards app. Additionally, the analysis examines the effects of students' activity on vocabulary learning. The analysis is based on data extracted from the app and focuses on 1) the percentage of students joining the app per week for each group and 2) the total number of days and minutes students spent learning from the app (study time) for each group. First a preliminary analysis is conducted for each followed by a mixed effects analysis for study time.

5.3.4.1 Preliminary analysis

Of the 106 students participating in this study, slightly more than half (54%, N = 58) signed up for the digital flashcard app. Figure 21 shows the percentage of students enrolling in the digital flashcard app by week. It shows that the four groups differed notably in the percentage and time of joining the app. On the pretest week, where students were given instructions on how to access the app, the largest percentage of involvement was for the weekly group (77%), followed by the biweekly group (55%), next was the monthly group (35%) while the no-quiz group where the least to join the app (14%).

Figure 21. The percentage of students joining the app by group and week



In week one, where the weekly group had their first quiz, the percentage of students joining the app increased slightly from 77% to 81%. Surprisingly, no increase in the number of new users was observed for the no-quiz group in week one even though their access was a classroom requirement (2% of course credit per week). A gradual increase in the percentage of involvement took place up to week three, followed by a plateau during the three weeks of school break where minor increases occurred. An increase in week four can be observed for the monthly group where they took their first quiz, the biweekly their second and the weekly group their fourth. Apart from a small increase in the weekly group, the involvement generally flattened out after week four until the end of the experiment.

In summary, more than half of the students joined the app during the first week for the weekly and biweekly groups while reaching the same percentage took three weeks for the monthly group and a whole month for the no-quiz group. By the end of the experiment, the total percentage of students learning from the app was 92% for the weekly group, 72% for the biweekly group and 57% for both the monthly and no-quiz groups.

Analyzing study time as measured by the number of days and minutes students in each group spent learning from the app provides some insights into how the experiment condition (quizzes vs. no quizzes and quiz frequency) along with individual differences (motivation and self-regulation) affect study time.

Table 23 shows the total number of days students accessed the digital flashcard app for each group. On average, students in the biweekly group spent more days learning from the app than the other groups with an average of 6.34. Students in the weekly group were second, learning for 5.83 days on average during the experiment. The no-quiz was second to last with an average of 3.92 learning days. Finally, the monthly group was the group with the least learning days with an average of 2.52. A more detailed overview is shown in Table 23.

Table 23. Descriptive statistics of the total number of days students spent learning from the flashcard app

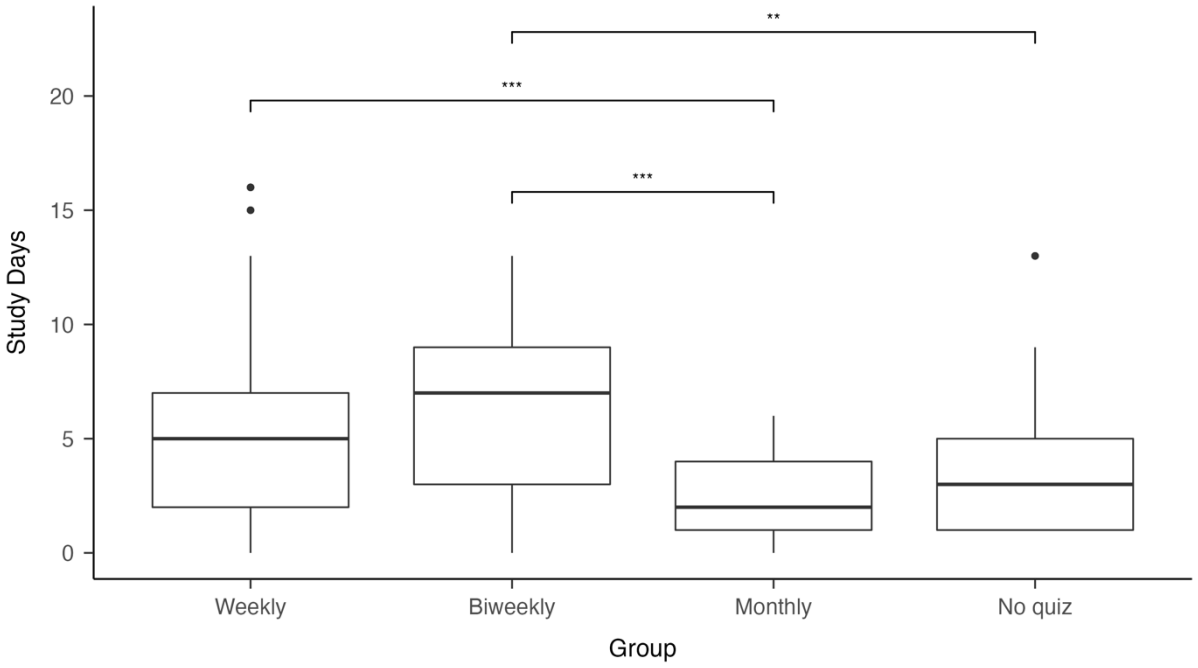
Group	Min	Max	Median	IQR	Mean	SD	SE	CI
Weekly	0	16	5	5	5.83	4.69	0.60	[4.62, 7.05]
Biweekly	0	13	7	6	6.34	3.92	0.59	[5.14, 7.56]
Monthly	0	6	2	3	2.52	1.92	0.31	[1.89, 3.16]
No-quiz	1	13	3	4	3.92	3.37	0.45	[3.02, 4.83]

Note. CI = 95% confidence interval

To test the significance of these differences, a Kruskal–Wallis test was used followed by post hoc tests. Results showed a significant main effect for group on the total number of study days ($H(3) = 25.82$, $p < .001$). Pairwise comparison (Figure 22) showed that the biweekly group had significantly higher study days than the monthly ($H = 57.02$, $p < .001$) and no-quiz ($H = 36.61$, $p = .008$) groups but not the weekly group ($H = 12.40$, $p = 1.000$). The weekly group had significantly higher levels of study days than the monthly group ($H = 44.62$, $p < .001$) but not

the no-quiz group (H = 19.41, p = 1.000). No significant difference existed between the monthly and the no-quiz groups (H = 20.41, p = .519).

Figure 22. Number of study days for each group with pairwise comparisons.



A more fine-grained measure of study time can be obtained by examining the minutes students spent learning on the app (Table 24). On average, students in the weekly group spent more minutes learning from the app than the other groups with an average of 21.46 across the whole term. Students in the biweekly group were second with an average of 12.79 minutes. The monthly group average was 8.37 learning minutes which was not markedly different from the no-quiz group average of 7.14 minutes. Overall, the total minutes of learning seem to increase as quiz frequency increases, although the results of a Kruskal–Wallis test showed no significant main effect for group on the total number of study minutes (H (3) = 5.69, p =.127).

Table 24. Descriptive statistics of the total number of minutes students spent learning from the flashcard app

Group	Min	Max	Median	IQR	Mean	SD	SE	CI
Weekly	0	98	8.5	37	21.46	29.47	3.80	[12.9, 29.1]
Biweekly	0	55	9	11	12.79	15.21	2.32	[8.11, 17.5]
Monthly	0	32	6	14.75	8.37	8.95	1.45	[5.43, 11.3]
No-quiz	0	22	7	4	7.14	5.47	0.73	[5.68, 8.61]

Note. CI = 95% confidence interval

5.3.4.2 Modeling the effect of study time on vocabulary learning

To examine whether study time as measured by days and minutes has a significant effect on meaning recognition and meaning recall vocabulary learning for each group, a series of mixed effects models were constructed. The mixed effects models here differ from the previous models in that they had the posttest as a dependent variable and the pretest as a control variable. This is because students' scores on the pretest are of less relevance as they did not start learning from the app at the time of the pretest. The models also had group as a fixed effect while subject and item were included as random effects. These were followed by second models with an interaction between study time and group to see if the effect of study time varies by group. The analysis was conducted on 44 students who took both the pretest and posttest and learned from the digital flashcard app.

For study time as measured by days, the model with study days showed improved model fit compared to the base model ($\chi2$ (1) = 4.34, p = .037). The model did not improve further with a two-way interaction between study days and group ($\chi2$ (3) = 3.57, p = .311). Results showed a significant and positive main effect for study days on meaning recognition test scores (b = 0.07, z = 2.14, p = .032). For meaning recall, adding the total number of study days to the

model did not improve model fit significantly without an interaction ($\chi 2$ (1) = 2.74, p = .097) or with an interaction with group ($\chi 2$ (4) = 7.92, p = .094).

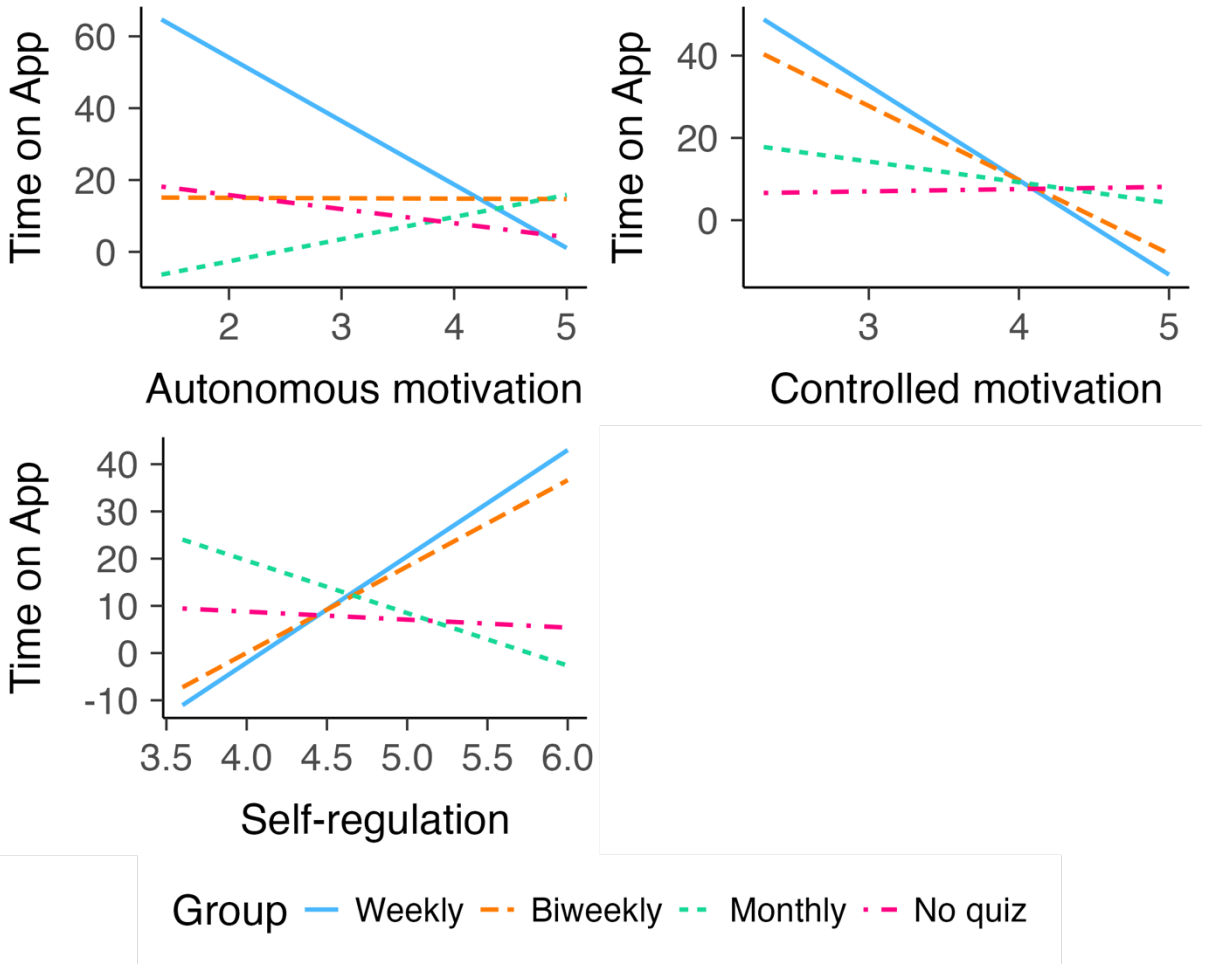
Similar results were found for study time as measured by minutes. For meaning recognition knowledge, the model with study minutes showed improved model fit compared to the base model ($\chi 2$ (1) = 8.00, p = .004). The model did not improve further with a two-way interaction between study minutes and group ($\chi 2$ (3) = 5.43, p = .142). Results showed a significant main effect for study minutes on meaning recognition test scores (b = 0.02, z = 2.95, p = .003). For meaning recall, adding the total number of study days to the model did not improve model fit significantly without an interaction ($\chi 2$ (1) = 1.04, p = .307) or with an interaction with group ($\chi 2$ (4) = 6.68, p = .153).

An additional analysis was conducted to examine the effect of motivation and self-regulation on the amount of time spent learning from the app. Due to convergence issues with mixed effects models, a multiple regression was constructed. Time on app was added as a dependent variable while autonomous motivation, controlled motivation and self-regulation were included as independent variables each in an interaction with group (i.e., two-way interactions). All the interactions were significant. Results of simple slope analysis (Figure 23) showed that autonomous motivation had significant positive effect on time on app for the monthly group (b = 6.17, t = 2.22, p =.028). In contrast, autonomous motivation had a significant negative effect on time on app for the weekly group (b = -17.67, t = -3.31, p =.001). This indicates that higher levels of autonomous motivation were associated with more time learning from the app for the monthly group but less learning for the weekly group. No significant effects were found for the biweekly and no-quiz groups (all ps >.05).

Controlled motivation had a significant negative effect on time on app for the weekly (b = -17.95, t = -3.17, p = <.002) and biweekly groups (b = -22.96, t = -6.12, p = <.001). These results

suggest that higher levels of controlled motivation were associated with spending less time on the app. No significant effects were found for the monthly and no-quiz groups (all ps >.05). Finally, self-regulation had a significant positive effect on time on app for the weekly (b = 22.51, t = 5.28, p = <.001) and biweekly groups (b = 18.28, t = 2.18, p =.030). suggesting that more self-regulated students spent more time learning from the app. Meanwhile, self-regulation had a significant negative effect on time on app for the monthly group (b = -11.12, t = -2.48, p =.014) suggesting that students in this group with higher levels of self-regulation spent less time on the app. No significant effects were found for the no-quiz group (p >.05).

Figure 23. The effect of motivation and self-regulation on time spent on app



In summary, the analysis overall showed that the groups differed significantly in the number of total days studied but not in the total number of minutes. Results also showed that increased

study time (as measured by both total study days and minutes) led to more significant vocabulary learning to the meaning recognition level but not meaning recall. Finally, the effects of motivation and self-regulation on app usage varied: autonomous motivation and self-regulation increased usage in some groups while decreasing it in others, with controlled motivation generally reducing usage.

5.3.5 Students' perceptions of quizzing and digital flashcards

The final part of the analysis focuses on analyzing students' perceptions and opinions regarding frequent quizzes and digital flashcard learning (the five scales were: quiz joy, quiz effectiveness, app joy, app effectiveness and future use). Similar to the previous section, the analysis starts by exploring the questionnaire and testing the significance of the differences among groups. This is followed by a second analysis which aims to explore the effects of these perceptions on students' vocabulary tests scores using mixed effects modeling. Similar to the app activity analysis, the models were constructed with the posttest as a dependent variable and the pretest as a control variable. The analysis was conducted on 57 students who completed the perceptions questionnaire. When analyzing students' perceptions of the app, 13 students were removed since the records showed that they did not access the app once (remaining students N = 44).

5.3.5.1 Preliminary analysis

Table 25 shows the descriptive statistics of students' responses to the perceptions questionnaire. The biweekly group expressed higher levels of enjoyment with frequent quizzes (3.88) than the weekly (3.68) and monthly groups (3.60). A Kruskal–Wallis test showed a significant main effect for group on students' responses on the quiz joy scale ($H(2) = 6.68$, $p = .035$). Post hoc analysis showed that the biweekly group enjoyed the quizzes significantly more than the monthly group ($H = 27.18$, $p = .031$) while all other pairwise comparisons were

not significant (all ps > .05). All three groups who received quizzes perceived them to be effective in vocabulary learning. The weekly group average was 3.99, the biweekly group average was 4.16 and the monthly average was 4.12. No significant differences existed among the three groups (H (2) = 3.30, p =.191).

Table 25. Descriptive statistics of the perceptions questionnaire scales

	Quiz joy		Quiz eff.		App joy		App eff.		Future	
	M	SD	M	SD	M	SD	M	SD	M	SD
Weekly	3.68	0.98	3.99	0.91	3.25	0.71	4.12	0.57	3.34	0.71
Biweekly	3.88	0.78	4.16	1.02	3.71	0.50	4.22	0.89	4.21	0.50
Monthly	3.60	0.82	4.12	0.70	2.83	0.52	3.76	0.85	3.28	0.52
No-quiz	-	-	-	-	3.07	0.93	3.83	0.85	3.24	0.93

Notes. eff. = effectiveness. Max score for all scales = 5.

Students' responses on the app enjoyment scale showed that the weekly (3.25) and biweekly (3.71) groups expressed higher levels of enjoyment with the app than the monthly (2.83) and the no-quiz groups (3.07). A Kruskal–Wallis test showed a significant main effect for group on students' responses on the app joy scale (H (3) = 24.33, p < .001). Post hoc analysis showed that the biweekly group enjoyed the app significantly more than the weekly (H = 31.07, p =.024), monthly (H = 59.96, p < .001) and the no-quiz group (H = 21.00, p =.001) while all other pairwise comparisons were not significant (all ps > .05).

Results of the app perceived effectiveness scale showed that both the weekly group (4.12) and the biweekly group (4.22) expressed higher levels of perceived effectiveness than the monthly (3.76) and no-quiz groups (3.83). A Kruskal–Wallis test showed a significant main effect for group on students' responses on the perceived app effectiveness scale (H (3) = 9.87, p = .019).

Post hoc analysis showed that the biweekly group perceived the app to be significantly more effective than the no-quiz group (H = 28.78, p =.044) while all other pairwise comparisons were not significant (all ps > .05)

Finally, the future scale, which investigates students' willingness to continue future vocabulary learning from the app, showed that the biweekly group was different from the other groups in showing a higher level of intention to continue learning from the app after the end of the experiment (4.21). The weekly group expressed slightly higher intentions to learn vocabulary from the app in the future (3.34) than the monthly (3.28) and the no-quiz groups (3.24). A Kruskal–Wallis test showed a significant main effect for group on students' responses on the quiz joy scale (H (3) = 16.26, p < .001). Post hoc analysis showed that the biweekly group expressed significantly more willingness to continue learning from the app than the weekly (H = 34.39, p =.007), monthly (H = 41.46, p =.007) and no-quiz groups (H = 38.16, p =.001) while all other pairwise comparisons were not significant (all ps > .05).

5.3.5.2 Modeling the effect of students' perceptions of frequent quizzes and digital flashcard learning on vocabulary learning

The analysis here is similar to the analysis in study time. The models included the posttest scores as a dependent variable with the pretest scores as a covariate and the five perceptions scales as fixed effects (quiz joy, quiz effectiveness, app joy, app effectiveness and future use). The scales were analyzed separately, once without interactions and second with interactions with group. The models had vocabulary test scores as dependent variables which assumes that they are predicted by students' perceptions. However, the relationship could also be perceived in the opposite direction in which vocabulary scores predict students' perceptions. In other words, it is possible that a more positive perceptions leads to higher test scores but it is also possible that higher test scores lead to a more positive perceptions. It is not possible to

determine the direction of effects in regression analysis (Field et al., 2012). The analysis here is conducted with the hypothesis that having more positive perceptions of a learning technique can have an effect on the learning outcome (see section 5.1.4)

Students' quiz enjoyment did not improve meaning recognition model fit significantly ($\chi2$ (1) = 1.03, p = .308) nor the model with an interaction with group ($\chi2$ (3) = 4.45, p = .216). The same for meaning recall, quiz joy levels did not have a significant effect on test scores in the model without an interaction with group ($\chi2$ (1) = 1.47, p = .225) or with an interaction ($\chi2$ (3) = 2.02, p = .566). Students' quiz perceived effectiveness did not improve meaning recognition model fit significantly ($\chi2$ (1) = 1.46, p = .226) nor the model with an interaction with group ($\chi2$ (2) = 1.86, p = .393). For meaning recall, quiz perceived effectiveness improved model fit significantly ($\chi2$ (1) = 6.18, p = .012), but not the model with an interaction with group ($\chi2$ (2) = 2.32, p = .312). Results showed that quiz perceived effectiveness had a significant main effect on meaning recall test scores (b = 1.52, z = 2.37, p = .017). Students who perceived the frequent quizzes to be effective had higher meaning recall vocabulary test scores regardless of which quiz frequency group they were in.

Perceived app enjoyment did not improve meaning recognition model fit significantly ($\chi2$ (1) = 0.27, p = .601) nor did the model with an interaction with group ($\chi2$ (4) = 7.22, p = .124) The same for meaning recall, the inclusion of perceived app enjoyment did not improve model fit significantly in the model without an interaction with group ($\chi2$ (1) = 2.20, p = .137) nor the model with an interaction ($\chi2$ (4) = 2.88, p = .577). Perceived app effectiveness ratings did not improve meaning recognition model fit significantly ($\chi2$ (1) = 0.10, p = .745) nor the model with an interaction with group ($\chi2$ (4) = 1.12, p = .890). The same for meaning recall, the inclusion of perceived app effectiveness did not improve model fit significantly in the model without an interaction with group ($\chi2$ (1) = 0.23 p = .625) or the model without an interaction ($\chi2$ (4) = 3.4, p = .493). Finally, Students' willingness to keep using the app in the future did

not improve meaning recognition model fit significantly ($\chi2$ (1) = 0.34, p = .554) nor the model with an interaction with group ($\chi2$ (4) = 4.11, p = .390). The same for meaning recall, the inclusion of perceived app effectiveness did not improve model fit significantly in the model without an interaction with group ($\chi2$ (1) = 0.00 p = .953) or the model with an interaction ($\chi2$ (4) = 6.63, p = .156).

Overall, the analysis showed that the biweekly group had the highest positive perceptions of both quizzes and digital flashcard learning followed by the weekly group. Students in general showed more agreement on the effectiveness of both quizzes and digital flashcards than on enjoyment. Only students' perceptions of the effectiveness of frequent quizzes had a significant positive effect on vocabulary test scores (meaning recall). Finally, students (except for the biweekly group) were neutral about continuing to learn vocabulary from the app in the future.

5.4 Discussion

The aim of the second study was to find an effective intervention for EFL students struggling with learning high frequency vocabulary. The results of the first study showed clear limitations in Saudi EFL learners' knowledge of high frequency words in that less than 1% of the students were able to master the first most frequent 1000 words. Flashcards offer an effective and efficient way of learning words and digital flashcards in particular can be applied out-of-class given that class time is limited in many EFL classes. However, studies have found that the majority of students do not engage in out-of-class vocabulary learning if it is optional, thus class credits were used in the past to motivate them to study (Seibert Hanson & Brown, 2019). Still, using class credit only as a means of getting students to study suffers from a number of limitations. First, students might access the digital flashcards just to get the credits and do not necessarily learn since they do not have to demonstrate their learning. Second, tracking who studied and who did not might be laborious and add more workload to the already busy

schedule of many EFL teachers. Results overall showed that combining digital flashcard learning with frequent in-class quizzes can be effective and can lead to better vocabulary learning than relying on class credit only.

5.4.1 The effect of quizzes

Based on the direct and indirect benefits of testing (Roediger et al., 2011), it was hypothesized that the quiz group would outperform the no-quiz group. Results confirmed this hypothesis and showed that pairing out-of-class flashcard learning with quizzes resulted in significant vocabulary learning to the meaning recognition (mean gain = 14.40) and meaning recall level (mean gain = 6.02). On the other hand, the no-quiz group did not learn vocabulary significantly by the end of the semester either to the meaning recognition (mean gain = -0.39) or meaning recall level (mean gain = 0.91). The main factor perhaps that led to the significant gains of the quiz group and the nonsignificant gains of the no-quiz is the testing effect (Karpicke & Roediger, 2007) . Taking quizzes seemed to enhance word retention for the quizzed group while no such advantage was available for the no-quiz group. One of the main reasons why taking a quiz improves retention is the fact that taking quizzes provides opportunities for retrieval practice. As discussed before 5.1.2, retrieval practice is the process of recalling previously stored information which has been shown to enhance learning (Barcroft, 2007). Students who received quizzes had the opportunity to retrieve the target words during quizzes which might have helped them remember these words on the vocabulary posttests and score higher than students who received no quizzes.

The finding that the no-quiz group made no significant gains is not in line with Seibert Hanson and Brown's study (2019) which found that class credits alone were enough to motivate students to study. Keeping in mind the fact that the present study allocated more class credit (16%) than the Seibert Hanson and Brown study (10%), this still was not enough to generate

significant vocabulary learning for the no-quiz group. This finding suggests that class credits alone are not enough to motivate students to learn vocabulary out-of-class from digital flashcards in all contexts and perhaps a more effective approach would be the inclusion of frequent in-class quizzes. The lack of significant gains might be due to the fact that the average study time per week for the students in this study was three times smaller than the average study time per week for the students in Seibert Hanson and Brown's study (2019). Results of the study time analysis showed that the average number of study days per week for the students in this study was 0.60 days per week while the average for the participants in the Seibert Hanson and Brown study was 1.72 days per week (SD = 1.93). Although the students in the no-quiz group reported that they were motivated, their motivation level might not have been strong enough to study vocabulary out-of-class to make significant gains. Overall, these findings suggest that supplementing out-of-class flashcard learning with in-class quizzes not only enhances learning, but in some cases might be necessary for learning.

5.4.2 The effect of quiz frequency

The second research question investigated the key question of how frequently quizzes should be given to students (i.e., weekly, biweekly or monthly). The weekly group had a number of potential advantages over the biweekly and monthly groups. First, they were quizzed on all 120 words, unlike the biweekly and monthly groups who were quizzed on a random sample of the target words (biweekly: 60, monthly 30) potentially enhancing the advantage gained from the testing effect. Relatedly is the fact that they had more in-class exposure to the target words given that they had eight quizzes in total which is twice as many as the biweekly group and four times as many as the monthly group.

Results however showed that the gains of the weekly (14.39) biweekly (13.93) and monthly groups (14.93) on meaning recognition vocabulary were very similar. Likewise, the gains of

the weekly (4.65) biweekly (9.63) and monthly groups (3.07) on meaning recall vocabulary were not significantly different from one another. These findings are in line with Bangert-Drowns' (1991) meta-analysis study on quiz frequency which found that having one quiz during a semester is effective but having more does not lead to more substantial learning.

Despite the potential advantages associated with more frequent quizzes, students in the weekly group did not learn significantly more vocabulary than groups with less frequent quizzes. One potential explanation for why the groups with less frequent quizzes made comparable vocabulary gains might be related to the whole and part learning distinction (Nakata & Webb, 2016). The whole learning of, for instance, 50 words would involve learning all of these words together while part learning would involve splitting these 50 words into for example five blocks each containing 10 words. Research suggests that whole learning is more effective than part learning due to longer spacing between when a word is met and the next repetition (the lag effect; Kornell, 2009; Nakata & Webb, 2016). The students in the monthly group had the advantage of longer lags (59 words until next repetition) while the weekly group had shorter lags (14 words until next repetition). Similarly, the biweekly group had longer lags (29 words until next repetition) than the weekly group. The lag effect could be perhaps what enabled the groups with less frequent quizzes to have vocabulary gains similar to the weekly group. The fact that three groups were not significantly different from one another might indicate that different quiz frequencies come with different advantages[11]. More frequent quizzes seem to amplify the testing effect in that there is a chance for more words to be tested while less frequent quizzes seem to have the advantage of within-session spacing or the lag effect (Kornell, 2009).

Another explanation might come from the fact that quizzes in this study were non-cumulative (i.e., previously tested items did not appear in subsequent quizzes). Thus, more frequent

[11] Assuming that only a sample of words are tested with less frequent quizzes and not all words.

quizzes in this case do not have spacing (i.e., distributed retrieval) advantage over less frequent quizzes. Gurung and Burns (2019, p. 739) state that "distributing study sessions across time does not constitute spaced practice if different material is studied in those sessions". No study appears to have investigated the effect of quiz frequency with cumulative quizzes. However, studies comparing cumulative to non-cumulative weekly quizzes show an advantage for cumulative testing (Nakata et al., 2021). Nakata et al. (2021) compared cumulative weekly quizzes to non-cumulative weekly quizzes. Both groups were required to learn 10 target words every week over a school semester. Also, both groups were quizzed receptively and productively on the words introduced on the previous week. However, only the cumulative group was quizzed on the words introduced on all previous weeks. Results of comprehensive posttests showed that cumulative quizzes were twice (receptive recall) and three times (productive recall) more effective than non-cumulative quizzes. When quizzes are cumulative, then more frequent quizzes might show an advantage over less frequent quizzes possibly due to increased opportunity for retrieval practice (i.e., studying the same word more, over distributed periods of time). The lack of increased distributed retrieval might be another explanation for the lack of advantage for more frequent quizzes in this study.

Results of the comparison between the quiz and no-quiz groups showed significant vocabulary gains on meaning recognition and meaning recall. When the quizzed groups (weekly, biweekly and monthly) were analyzed separately, the effect of quizzing was limited to meaning recognition knowledge but not meaning recall. This could be explained in terms of transfer-appropriate processing theory (Morris et al., 1977), which posits that memory performance is enhanced when the cognitive processes used during learning match those required during testing, as the weekly quizzes were administered in a meaning recognition format. Additionally, this might be due to students in all groups not spending much time learning from the app to the level that results in significant and deeper learning to the meaning recall level.

Despite the seemingly low learning time from the app, students in the quizzed groups managed to learn words significantly to the meaning recognition level. This suggests that digital flashcard learning paired with frequent in-class quizzes can be very helpful in forming initial associations between words and their meanings which could be deepened later through more retrievals and various encounters (Webb & Nation, 2017).

5.4.3 The role of individual differences

The aim of the third research question was to examine the effects of autonomous motivation, controlled motivation and self-regulation on digital flashcard learning in out-of-class settings. One of the main purposes of including a no-quiz group was to test whether relying on students' motivation levels and self-regulation skills alone would lead to significant vocabulary learning without the need for frequent quizzes. The results from the first research question clearly showed that this was not the case. There is yet a possibility that the lack of significant vocabulary gains was not due to a lack of quizzes but perhaps that the no-quiz group had lower motivation and self-regulation levels than the other three quizzed groups. The following discussion focuses only on the results of the best-fit comprehensive models.

The four groups differed significantly in their autonomous motivation levels. Although the biweekly group was significantly more autonomously motivated than the weekly and monthly groups, they were not more autonomously motivated than the no-quiz group. Results of the comprehensive models showed that autonomous motivation had a positive and significant main effect on meaning recall test scores but not meaning recognition. Results also showed that the gains the quizzed groups made remained significant even after controlling for autonomous motivation. The fact that the no-quiz group did not demonstrate significantly lower autonomous motivation levels than the other three groups, and that the gains of quizzed groups remained significant while controlling for autonomous motivation, suggests that the learning gains are

robust and not heavily dependent on students' autonomous motivation. In other words, the inclusion of quizzes seems to lead to significant gains regardless of students' autonomous motivation. Based on this study findings, being autonomously motivated alone did not lead to significant vocabulary learning while being quizzed frequently led to significant vocabulary learning. Autonomous motivation alone might not be enough for vocabulary learning in out-of-class settings, and students might need an additional incentive to learn which in-class quizzes appear to offer. This might be due to frequent quizzes providing students with a more influential drive to learn than autonomous motivation. Another possibility is that even motivated students might find it difficult to maintain their language learning motivation for an extended period of time especially in an out-of-class context (García Botero et al., 2019). Frequent and regular quizzes seem to help students be more consistent in their out-of-class learning.

Similarly, the possibility that the lack of significant gains for the no-quiz group was because they were possibly mainly driven by controlled motivation (which tends to lead to lower language learning outcomes; Alamer, 2021a) was not supported by the study findings either. The controlled motivation scale results showed that the no-quiz group did not demonstrate significantly higher levels of controlled motivation than the other three groups. In fact, the results of both meaning recognition and meaning recall models showed that controlled motivation had a significant negative effect only on the weekly and biweekly groups. These negative effects however did not prevent students in these groups from learning vocabulary significantly, nor did they lead to lower gains compared to the monthly group. This finding, coupled with the autonomous motivation findings, suggests that the effect of motivation seemed to be overshadowed by the testing effect. Again, it seems that what mattered most was not students' motivation levels but whether they were in a class that had quizzes. Similar findings were reached in Seibert Hanson and Brown (2019) who found that using for grades

digital flashcard learning in an out-of-class context led to significant vocabulary learning even after controlling for motivation. Overall, the findings suggest that digital flashcard learning paired with in-class quizzes can be very effective and robust, displaying minimal sensitivity to variations in students' motivation levels.

Of course, the above discussion does not imply that the effect of motivation should be ruled out entirely, but points out that higher levels of autonomous motivation alone did not compensate for the absence of quizzes. Having said that, motivation did affect the amount of learning. Students with more autonomous motivation learned more vocabulary to the meaning recall level regardless of group. The finding is consistent with study one results and with research on vocabulary that higher levels of autonomous motivation often lead to more vocabulary learning (Alamer, 2018; J. H. Lee et al., 2022; Y. Zhang et al., 2017). In contrast, controlled motivation had negative effects on the scores of the weekly and biweekly groups while no significant effects were found for the monthly and the no-quiz groups. This finding is in agreement with previous studies which mostly found either negative or no effect for controlled motivation on language learning outcomes (Alamer, 2021a; Liu, 2007; Noels et al., 1999; F. X. Wang, 2008), including study one which found no significant effects. The findings of both autonomous and controlled motivation in this study lend general support to the self-determination theory conceptualization of motivation (Ryan & Deci, 2017). In this study, students who learn English because it is intrinsically rewarding or desirable managed to learn vocabulary significantly more to the meaning recall level. On the other hand, students who reported learning English for external reasons such as guilt or grades learned significantly lower vocabulary on both meaning recognition and meaning recall. Overall, students who learn English because they want to (autonomous motivation) learned more words than students who learn English because they have to (controlled motivation).

Self-regulation was also investigated in this study to examine its effect of vocabulary learning from digital flashcards in out-of-class settings. The levels of self-regulation of the four groups were compared using non-parametric ANOVA and the results showed no significant differences between the groups. The fact that self-regulation was similar across all groups also rules out the possibility that the gains made by the quiz groups were because they had better self-regulation skills than the no-quiz group. Results of comprehensive mixed effects models showed no significant effect for self-regulation on meaning recognition or meaning recall knowledge in the comprehensive models. The findings also support the results of the first study, which found no significant effects for self-regulation on meaning recognition or meaning recall in the comprehensive models. Thus, the present finding serves as a replication for study one as it was conducted with students from the same school and grade level. As suggested in study one, the instrument might require adaptation to the Saudi context before providing meaningful insights into the self-regulation strategies employed by Saudi EFL students.

Overall, the investigation suggests that the integration of digital flashcard learning with in-class quizzes can lead to significant vocabulary gains that appear to be robust against individual variations in motivation and self-regulation.

5.4.4 Study time

One of the advantages of using digital flashcards in this study is that students' learning behavior can be tracked. Two key pieces of information were extracted from the app, namely the first time students accessed the app and the total amount of time spent learning (in days and minutes; see section 5.2.3). Results of when each group accessed the app for the first time (Figure 21) showed that the majority of the students in the weekly and biweekly groups joined the app in the first week. In contrast, it took a full month for half of the students in the no-quiz group to

initiate learning from the app. This suggests that more frequent quizzes might lead to earlier engagement with vocabulary learning in out-of-class settings.

By the end of the experiment, more than 40% of the students in the monthly and no-quiz group did not access the app once throughout the semester. In contrast, the groups that had more frequent quizzes demonstrated a higher rate of involvement in learning from the app, with only 8% of students not registering in the app in the weekly group and 28% in the biweekly group. This finding indicates that higher quiz frequency seems to correspond to higher involvement rates yet this did not lead to more significant gains.

On a more individual level, students in this study varied in the amount of time spent learning from the app with some spending up to 98 minutes throughout the semester while others spending only few minutes. Motivation and self-regulation were the two sources of individual variation that were examined in this study. The analysis showed that students in the weekly and biweekly groups with higher levels of self-regulation spent more time on the app. This fits with framework of self-regulated learning which suggest that more self-regulated learners are more effective in their learning (e.g., planning learning and avoiding distraction, see section 2.3.2). In terms of motivation, higher levels of autonomous motivation were associated with spending more time on the app for the monthly group. This finding is reasonable since students driven by autonomous motivation tend to find pleasure in language learning and are more likely to spend more effort during learning (Alamer, 2021a). In contrast, results showed that students with high controlled motivation spent less time on the app. Students with higher levels of controlled motivation tend to spend less time studying materials (Kusurkar et al., 2013) possibly due to lack of interest and enjoyment. What was rather unexpected, was that higher levels of autonomous motivation and self-regulation had negative effects on the amount of study time for the weekly group and monthly group respectively. As discussed above, these factors are usually associated with more effective learning. However, it is possible for

motivated students to lack engagement (i.e., translating motivation to action) due to factors such as the challenge of language learning and competing priorities in learners' lives (Hiver et al., 2020; Teravainen-Goff, 2022). Similarly, some students might benefit from more training in self-regulation to be more effective learners especially in out-of-class settings where they are more responsible for their own learning (García Botero et al., 2019). These factors might explain the low study time of some motivated and self-regulated students in this study.

The discussion of the first research question explained the advantage of the quizzed groups only in terms of the direct effects of testing (i.e., the testing effect). However, research points out to the indirect effects of testing (most notably increased study time) as a possible additional explanation for the outperformance of tested groups. The use of digital flashcards allows us to separate the direct effect of testing (i.e., retrieval practice) from a chief indirect effect of testing (i.e., increased study time) in naturalistic settings and check such a conclusion. In particular, the analysis of students' activity on the app and the total number of days and minutes students spent learning provide useful insights. The results showed that despite the fact that the weekly and biweekly groups' total number of days spent learning was significantly larger than the no-quiz group, the monthly group had fewer study days than the no-quiz group. That is, although students in the monthly group studied fewer days on the app than students in the no-quiz group, they scored significantly higher on the meaning recognition posttest. The monthly group vocabulary gains were also not significantly different from the weekly and biweekly groups who spent more time learning from the app than the monthly group. Additionally, the study time as measured by minutes showed no significant differences between all groups. These findings combined suggest that the direct effects of testing (i.e., retrieval practice) might have been the key driving force in the testing effect and not the indirect effects of testing (i.e., increased study time) in this study. This conclusion receives support from numerous experimental studies which found significant learning resulting from the testing effect after

controlling for time (e.g., Robey, 2019). It should be noted, however, that some students might have learned the target words outside of the app, resulting in less study time, which in turn might have downplayed the influence of study time on vocabulary gains.

Mixed effects models were fitted to examine the effects of study time (as measured by days and minutes) on meaning recognition and meaning recall vocabulary. Results showed positive significant effects for study time as measured by both study days and minutes on meaning recognition vocabulary knowledge but not meaning recall. The finding that the more time students spend learning from a flashcard app the more vocabulary is learned is reasonable according to the crude principle of time on task (i.e., the more time spent on a task the more learning happening) and has also been found in other studies (Seibert Hanson & Brown, 2019). The lack of significant effect for study time on meaning recall vocabulary knowledge might be due to the fact that learning vocabulary to the meaning recall level is more difficult than learning vocabulary to the meaning recognition level (Laufer & Goldstein, 2004). Additionally, as suggested earlier, students in the present study did not overall spend much time learning from the app, which could have resulted in the nonsignificant effect on meaning recall test scores.

In summary, it appears that the primary driver of the advantage seen in the quizzed group is the direct effect of testing, rather than the indirect effect of increased study time. Additionally, it seems that the more time students spend on the flashcard app the more vocabulary they learn to the meaning recognition level.

5.4.5 Students' perceptions of quizzes and digital flashcards

At the end of the experiment, students were given a questionnaire to measure their perceptions of frequent quizzes and digital flashcard learning. The questionnaire focused on whether students perceived quizzes and digital flashcard learning as being effective and enjoyable.

Quizzes had two scales (perceived effectiveness and enjoyment) while digital flashcard learning from the app had three scales (perceived effectiveness, enjoyment and future use).

All three quizzed groups seemed to enjoy frequent quizzes to some degree. The biweekly group showed significantly higher levels of enjoyment compared to the monthly group but not the weekly group. This might be due to the biweekly group reporting higher autonomous motivation than the other groups and therefore being more positive toward quizzes in general. Similarly, all groups tended to perceive frequent quizzes as being effective for vocabulary learning without significant differences. There seems to be more agreement on quiz effectiveness than on quiz enjoyment. Overall, the findings are in line with previous studies which found that quizzes were generally perceived positively by learners (Bangert-Drowns et al., 1991; Deck, 1998; Kika et al., 1992).

Students were generally neutral regarding app enjoyment. Although the majority of studies have found positive perceptions of digital flashcard learning (Davie & Hilber, 2015; Sage et al., 2019, 2020), a small number of studies found negative perceptions (Seibert Hanson & Brown, 2019). The students in one study (Sage et al., 2020) who had positive perceptions of learning from flashcards described a digital flashcard app (Quizlet) as being fun, useful, helpful and convenient. In contrast, students in another study (Seibert Hanson & Brown, 2019) who had negative perceptions described a digital flashcard app (Anki) as being unengaging and basic. The students in the present study seem to sit somewhere in between these studies but lean more towards having a positive perceptions. One possibility for the lack of a negative perceptions might be because the interface of the app used here (Brainscape) appears to be more engaging than Anki and resembles Quizlet in design. At the same time, the lack of a clear positive perceptions aligns with the bitter pill view of digital flashcard learning in the sense that students in this study saw learning to be effective but not particularly enjoyable. The lack of a clear positive perceptions for digital flashcard learning might be attributed to the general

repetitive and monotonous nature of learning tens of words. The two qualities of monotony and repetitiveness have been linked to boredom in second language research (Kruk et al., 2021) but their negative effects might be offset by perceived effectiveness. Students in all groups tended to agree that the app was effective for vocabulary learning without much variation. Finally, students were asked directly about their willingness to continue learning from the app in the future. Results showed that all groups (except for the biweekly group) were neutral regarding continued app learning. This aligns with their neutral assessment of how much they enjoyed using the app. Only the biweekly group showed a significantly higher willingness to keep learning from the app in the future. This could be attributed to the fact that they showed a higher level of app enjoyment than the other groups.

Overall, there was more agreement among students on the effectiveness of both frequent quizzes and digital flashcard learning but less agreement on quiz enjoyment and even less agreement on app enjoyment. In addition, all groups except for the biweekly group were neutral about future learning from the app.

Results of mixed effects models aimed to examine the effect of students' perceptions of quizzes and digital flashcard learning on meaning recognition and meaning recall vocabulary learning. The five scales (quiz joy, quiz effectiveness, app joy, app effectiveness and future use) were included as fixed effects each in a separate model. Results showed no significant effect for quiz joy on meaning recognition or meaning recall vocabulary. Similarly, there was no significant effect for app enjoyment on meaning recognition or meaning recall tests. Quiz perceived effectiveness had a significant positive effect on meaning recall test scores regardless of group. This finding indicates that students who held positive perceptions of the effectiveness of frequent quizzes seem to learn more vocabulary to the meaning recall level. As suggested before, this can be seen the other way around in which performing well on the quizzes might have led students to develop a more positive perceptions of frequent quizzes. Finally, both

students' perceived app effectiveness and willingness to continue using the app in the future did not significantly predict meaning recognition or meaning recall test scores. Overall, results of mixed effects models seem to suggest that students' perceptions regarding the effectiveness of a learning technique seems to have more impact on vocabulary learning than their perceptions regarding its enjoyment. The implications of this and other findings on pedagogy are discussed in the next section.

5.5 Pedagogical implications

Flashcard learning is possibly the most efficient way of learning vocabulary (Webb et al., 2020). However, language learners seem to have issues initiating and maintaining learning from digital flashcards (Seibert Hanson & Brown, 2019). Thus, it is perhaps useful that flashcard learning is guided by the teachers in the early stages (Nation, 2022). One effective way of doing this based on the study findings is by introducing digital flashcards as an out-of-class activity with in-class quizzes. The current study presents a number of pedagogical implications that might improve students' vocabulary learning and learning from digital flashcards.

The first finding showed that class credits alone were not enough to encourage students to learn vocabulary out-of-class. What this means for teachers is that relying on class credits alone might not be sufficient and an additional incentive might be needed. This additional incentive can take the form of in-class quizzes, given that the groups who had quizzes in this study learned vocabulary significantly while no significant gains were observed in the no-quiz group.

Another useful finding is that supplementing digital flashcard learning even with one quiz a month can lead to significant vocabulary gains. The gains from taking one quiz a month were similar to the gains resulting from taking a quiz every week or every two weeks. This finding suggests no added value for more frequent quizzes. Based on this study, it seems that teachers

can opt for lower frequency quizzes without missing out on significant vocabulary gains. More frequent quizzes add extra workload to the already busy schedule of many language teachers. Therefore, it might be more practical to give students fewer quizzes as they appear to result in similar gains to more frequent quizzes.

Lack of motivation (or demotivation) is a major issue in many language learning classrooms (Tanaka, 2017). Similarly, students might not be readily able to self-regulate their learning and might struggle even more in out-of-class learning with technology (García Botero et al., 2021). The present study provided some positive results and showed that students made significant vocabulary gains even after controlling for motivation and self-regulation. This suggests that combining out-of-class digital flashcards with in-class quizzes can result in robust vocabulary learning that does not seem to be thwarted by students' motivation or self-regulation levels.

Although motivation did not undermine the learning gains, it had effects on the amount of learning. Students' gains were positively and negatively affected by their motivation levels. Thus, vocabulary learning from flashcards is maximized when students are autonomously motivated and minimized when they are driven by controlled motivation. Fostering autonomous motivation among students can lead to more optimal results when implementing digital flashcards as an out-of-class activity. Within the framework of SDT, autonomous motivation needs to be catered for by satisfying students' basic psychological needs (BPN) which are "innate psychological nutriments that are essential for ongoing psychological growth, integrity, and well-being" (Deci & Ryan, 2000, p. 229). The three components of BPN are autonomy, competence and relatedness (Deci & Ryan, 2000; Ryan & Deci, 2017). Autonomy refers to the feeling of volition and being in control in pursuing tasks that are personally meaningful. The term competence is used to refer to the feeling of being capable in carrying out tasks successfully and effectively. Relatedness refers to the feeling of belonging and being part of a community. Within the context of vocabulary learning from digital

flashcards, students' need for autonomy can be nourished by for example allowing them to choose which words they wish to learn to make their learning more personalized (while making sure they are aware of the importance of high and mid-frequency words). Students' sense of competence can be developed by for example starting with a small number of words to learn at the beginning and then they can work their way up by making incremental increases. Finally, students' sense of relatedness can be fostered by for example allowing peer assessment in which students grade the quizzes of their classmates leading to more collaborative learning.

Students are not passive recipients of information and their perceptions and reactions towards instructional techniques matter (Mantle-Bromley, 1995). The study found that students who had a positive perception regarding the effectiveness of frequent quizzes learned more vocabulary. Therefore, informing students about the benefits and effectiveness of frequent quizzes might enhance their vocabulary learning.

Although not an implication directly from this study, it is worth pointing out that while learning from flashcards is an effective and efficient way of learning vocabulary, it is important to remember that it should be part of a balanced language learning program (Nation, 2007; Webb & Nation, 2017). This involves creating opportunities for learners to learn vocabulary both intentionally (e.g., flashcards or vocabulary learning activities) and incidentally (e.g., reading and listening).

5.6 Conclusion

As stated in the introduction, the aim of the current intervention study was to help language learners improve their vocabulary learning. Classroom time for language instruction is usually limited to a few hours a week. It was hypothesized that supplementing out-of-class flashcard learning with in-class quizzes will lead to more effective vocabulary learning. The first finding showed that the group who had no quizzes did not learn vocabulary significantly from the

pretest to the posttest. In contrast, the groups who had quizzes learned vocabulary significantly to the meaning recognition level. This finding indicates that quizzes not only enhance learning but might be necessary for learning from digital flashcards to be successful in some contexts. Another goal of the study was to investigate the effect of quiz frequency (weekly, biweekly and monthly) on vocabulary learning. Results showed no significant differences in the learning gains between the three groups, suggesting that more frequent quizzes do not necessarily lead to more vocabulary learning. The gains that students in the quiz groups made were robust against individual differences. The quiz groups' gains were significant even after controlling for motivation and self-regulation. At the same time, motivation had significant effects on the amount of gains. More autonomously motivated students learned more vocabulary to the meaning recall level while students driven by controlled motivation learned less vocabulary to both meaning recognition and meaning recall levels. The amount of time that students spent learning from the app had a significant positive effect on meaning recognition vocabulary learning. Finally, students' positive perceptions of the effectiveness of quizzes had a positive effect on meaning recall vocabulary learning.

6. General discussion and conclusion

The overarching goal of the thesis was to investigate how vocabulary develops in a foreign language learning context and how this development can be improved. The thesis consists of two studies. Study one investigated vocabulary growth in an EFL context and examined how individual differences affect learning. Study two aimed to find ways to improve vocabulary learning in an EFL context. The first study helped in identifying that the majority of Saudi EFL students were having difficulties learning the highest frequency vocabulary (most frequent 1000 words). This was followed by the second study which focused on boosting their knowledge of high frequency words through intentional vocabulary learning using digital flashcards and frequent quizzes. The chapter begins by summarizing and discussing the main findings from the two studies. This is followed by a discussion of the overall implications, limitations and suggestions for future research and a conclusion.

6.1 Discussion of the main thesis findings

The aim of the first study was to examine how meaning recognition and meaning recall vocabulary knowledge develop over a school semester (12 weeks) and examine the role of individual differences, specifically out-of-class exposure, self-regulation and motivation in vocabulary learning. The participants were Saudi intermediate (16 years old) and secondary (17 years old) EFL learners. Results of meaning recognition tests showed that the intermediate students learned approximately 309 word-families (a significant gain) while the secondary students' gains of 99 word-families were not significant. Meaning recall gains were significant but lower for both

groups, where the intermediate students learned approximately 46 word-families while the secondary students learned around 82 word-families.

One way of evaluating the vocabulary growth of Saudi learners is by comparing their performance against a mastery level. It is reported that the Saudi Ministry of Education expects students to have a vocabulary size of around 3000 English words by the time they finish secondary education (Al-Masrai & Milton, 2012). This threshold seems reasonable given that previous studies showed that students can understand daily discourse if they know the most frequent 3000 word-families (van Zeeland & Schmitt, 2013). The secondary students in this study finished the first year knowing 979 word-families to the meaning recognition level. If they continue with the same rate of learning of 99 word-families per semester (99 x 3 semesters = potentially 397 words a year), they might finish secondary education with a vocabulary knowledge of around 1773 word-families which is far below the objective of the Ministry of Education. Similarly, although the intermediate students' meaning recognition growth was three times larger than that of the secondary students, their growth might decrease to the same level of the secondary students when they go to secondary education. One reason for this may be because textbooks in Saudi Arabia seem to introduce many words during intermediate years but then introduce fewer words during secondary education. The textbooks of secondary education seem to focus mainly on recycling previously introduced words which appears to come at the expense of learning new words based on study one findings. Additionally, after more than six years of instruction, only one student was able to master the first 1000 frequency band, and none mastered the remaining bands. Taken together, these findings suggest that vocabulary growth in an EFL context can be low and slow (Siyanova-Chanturia & Webb, 2016), and that students after many years of school instruction might not even develop good knowledge of the highest frequency band (i.e., most frequent 1000 word-families).

The low vocabulary knowledge of Saudi EFL learners is primarily due to the fact that Saudi students (like many EFL students globally) typically have limited exposure to English both in and out-of-class (Milton & Meara, 1998; Nurweni & Read, 1999; Siyanova-Chanturia & Webb, 2016; Webb & Chang, 2012). Additionally, although students in study one reported high levels of autonomous motivation, this might not necessarily translate to actions (i.e., engagement; Hiver et al., 2020). Teravainen-Goff (2022) notes that even some motivated students might fail to engage with language learning both in and out-of-class. To explain why, he surveyed 39 learners and teachers in England and Finland. Results revealed a number of reasons including disengaging classroom tasks and activities, the challenge of language learning and competing priorities in learners' lives. Although Saudi EFL learners reported higher levels of motivation, they may not engage or seek vocabulary learning opportunities perhaps due to the same reasons reported in Teravainen-Goff's study. Teachers also play a key role in students' vocabulary learning (e.g., planning how words are introduced and recycled; Webb & Nation, 2017). Studies on teachers in Saudi Arabia show that pre-service (Al-Masrai & Milton, 2012) and in-service English language teachers appear to have limited vocabulary knowledge (Alfairouz, 2015). Both studies report a vocabulary size of 5000 word-families and under, which is below the vocabulary size needed for unassisted language use (6000-7000 word-families for listening and 8000-9000 word-families for reading; Nation, 2006). Teachers also seem to have limited awareness of effective vocabulary instructional practices such as spacing word retrievals and the use of L1 translation (Sonbul et al., 2022).

The second study aimed to address the issue of limited vocabulary knowledge of Saudi EFL students and help them boost their knowledge of high frequency vocabulary. A meta-analysis found that flashcard learning was the most effective form of intentional vocabulary learning as

measured by effect size (Webb et al., 2020). It was assigned as an out-of-class activity given the limited time of many foreign language classes where students were asked to learn vocabulary from a digital flashcards app on their personal phones. To make sure that students actually learned from the app, in-class quizzes were employed. However, it was unclear based on the available research how frequently quizzes should occur for optimal out-of-class vocabulary learning. The second study aimed to address this gap by first examining the effect of quizzing (quiz vs. no-quiz) followed by an examination of the effect of quiz frequency (weekly, biweekly and monthly) on meaning recognition and meaning recall vocabulary learning over a school semester (12 weeks). It also examined the effect of motivation, self-regulation and perceptions on vocabulary learning. The first key finding was that the groups that were given frequent in-class quizzes learned vocabulary significantly to the meaning recognition level on the posttest while the group who did not receive quizzes did not make any significant vocabulary gains. The second key finding was that all three quiz frequency groups (weekly, biweekly and monthly) made similar vocabulary learning gains without any significant differences. Finally, all four groups (weekly, biweekly, monthly and no-quiz) did not learn vocabulary significantly to the meaning recall level by the end of the semester.

The first finding shows that students' willingness to engage in out-of-class language learning (i.e., extra-curricular learning see section 2.3.1) should not be taken for granted (Seibert Hanson & Brown, 2019). Although the students in the no-quiz group were assigned to learn from the app for course credit (16% - equivalent to the course credit earned by groups who took regular quizzes), this was still not enough to get them to learn effectively from the app to the level where they made significant vocabulary gains. In contrast, the quiz groups were able to learn vocabulary significantly by the end of the semester, possibly due to the testing effect (Roediger & Karpicke,

2006). This broadly suggests that for out-of-class language learning to be effective, employing tools such as quizzes may be needed to ensure students genuinely engage with the assigned materials.

The effect of quiz frequency on vocabulary learning has not been explored much in SLA. The studies in psychology are also not conclusive as to whether more quizzes lead to more learning (Bangert-Drowns et al., 1991; Beaulieu & Zar, 1986; Dustin, 1971; Keys, 1934; Palmer, 1974; Ross & Henry, 1939; Yang et al., 2021). The findings in this study showed that all three quiz frequency groups (weekly, biweekly and monthly) learned vocabulary similarly, suggesting that increasing quiz frequency does not necessarily increase vocabulary learning. This finding is in line with Bangert-Drowns' (1991) meta-analysis on quiz frequency, which found that having one quiz during a semester is effective but having more does not lead to more substantial learning. The weekly group had a number of potential advantages over the biweekly and monthly groups. First, they were quizzed on all 120 words over the course of the semester, unlike the biweekly and monthly groups who have been quizzed on a random sample of the target words (biweekly: 60, monthly 30) which might be expected to boost the advantage gained from the testing effect. Related is the fact that they had more in-class exposure to the target words given that they had eight quizzes in total which is twice as many as the biweekly group and four times as many as the monthly group. However, offering fewer quizzes for the same set of words seems to have a mechanism that compensates for the lack of these advantages. As suggested in the second study, the groups with fewer quizzes seemed to have benefitted from within-session spacing (i.e., lag effect; Kornell, 2009; Nakata & Webb, 2016). Where the weekly group was required to study 15 words before every quiz, the biweekly group was required to study 30 words and the monthly group 60 words. This increase in block size (which increases within-session spacing) has been

found to enhance vocabulary learning (Kornell, 2009; Nakata & Webb, 2016). Based on the results of study 2, the lower opportunities for retrieval practice and less exposure time associated with less frequent quizzes seem to be offset by the lag effect.

Another explanation might be related to the fact that the quizzes in this study were non-cumulative (i.e., previously tested words were not included in subsequent quizzes). Thus, more frequent quizzes did not have a spacing advantage over less frequent quizzes. As suggested by Gurung and Burns (2019), distributing study sessions over time does not qualify as spaced practice if different materials are studied during those sessions. The spacing effect, which can enhance vocabulary learning and retention (Nakata, 2008, 2020), was not greater with more frequent quizzes. The absence of enhanced distributed retrieval could be another reason why more frequent quizzes did not have an advantage in this study.

6.1.1 Meaning recognition and meaning recall growth

Examining the growth of recognition and recall vocabulary is central to understanding how vocabulary knowledge develops over time given that the two appear to be two distinct constructs and develop differently (González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004; Stewart et al., 2024). The first study is one of the few studies that examined how both meaning recognition and meaning recall vocabulary knowledge of the form-meaning link develops over time (Dóczi & Kormos, 2015). As expected, the findings showed that meaning recall vocabulary knowledge of the form-meaning link meaning lags behind meaning recognition vocabulary. The combined meaning recall vocabulary growth of both intermediate and secondary students (128 word-families) was three times lower than their meaning recognition growth (407 word-families). This is likely due to recall knowledge being more difficult than recognition (Laufer & Goldstein, 2004). One common explanation for the difficulty associated with recall comes from the two-stage theory

(J. Brown, 1976; Mandler, 2008). According to this theory, the process of recall involves two stages, starting with a search process followed by a decision process. What makes recognition easier is that it only involves the second stage (i.e., the decision process) and not the first. Recall is thus more prone to error or failure given that it involves more complex processes compared to the simpler process of recognition (Lachman & Forsberg, 1981). In a critical review of previous research, Mandler (2008) argues that the two-stage model (also referred to as the dual-process model) still provides a satisfactory explanation for the differences in recall and recognition memory.

In terms of vocabulary research, the finding that meaning recall growth lags behind meaning recognition is in line with González-Fernández and Schmitt's study (2019) where recognition and recall knowledge were found to be fundamentally distinct vocabulary constructs. In their cross-sectional study, González-Fernández and Schmitt examined how recognition and recall knowledge of the form-meaning link, derivations, multiple meanings and collocations develop in 144 Spanish EFL learners with varying levels of proficiency. They used implicational scaling analysis to identify if some aspects of vocabulary knowledge (e.g., form-meaning) are acquired before others (e.g., knowledge of multiple meanings). SEM was also used to examine the relationship between the word knowledge components. Results of the implicational scaling showed that recognition knowledge was easier to acquire than recall across all four tested aspects of vocabulary knowledge as shown below (easier to more difficult; González-Fernández and Schmitt, 2019, p.13):

Form–Meaning link meaning recognition > Collocate form recognition > Multiple-Meanings meaning recognition > Derivative form recognition > Collocate form recall > Form–Meaning link form recall > Derivative form recall > Multiple-Meanings recall

Additionally, the best-fit SEM model was a unidimensional construct of vocabulary, where recall and recognition were two different components loading onto the general vocabulary dimension. The findings of the first study provide support for this implicational scaling using longitudinal data in that students' meaning recognition vocabulary growth was larger than their meaning recall vocabulary growth (i.e., meaning recognition is easier than meaning recall). By showing that meaning recall growth lags behind meaning recognition due to learning to the meaning recall level being more difficult than meaning recognition learning, the first study is one of the few studies that have used longitudinal data to give support for the implication scaling in the form-meaning link aspect of vocabulary knowledge which has been proposed before based on cross-sectional data (Dóczi & Kormos, 2015; González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004).

The closest (and seemingly only) longitudinal study that compares meaning recognition and meaning recall knowledge is a study by Ozturk (2012) who compared receptive (meaning recognition) and productive (form recall) growth. She used both cross-sectional data (n = 55) and longitudinal data (n=17) with a period of three years between the two tests to measure the vocabulary growth of advanced EFL learners in a university in Turkey. The study used the VLT (Schmitt et al., 2001) and the Productive Levels Test (Laufer & Nation, 1999). One issue when comparing receptive and productive knowledge using different test formats (meaning recognition

and meaning recall) is that the findings might be biased toward receptive knowledge since recognition tests are easier than recall tests (Webb, 2005). A better approach to avoid this bias is the use of identical test formats (e.g., receptive recall vs. productive recall or receptive recognition vs. productive recognition) when comparing receptive and productive vocabulary knowledge (Webb, 2005). Also, the tests used have different words which might introduce uncontrolled variation due to word-related differences. The use of identical target words eliminates this unwanted variation. With these limitations in mind, the results showed no significant gains on the receptive tests on both the cross-sectional and longitudinal data. In contrast, students made significant gains on the recall test (10% increase). Ozturk explained this by suggesting that the participants were advanced (receptive test scores generally ranged from 24-28 out of 30 in the high frequency bands) which potentially left little room for improvement to be detected (i.e., a ceiling effect, see section 2.2.1 for potential issues in vocabulary growth assessment).

Relevant to the recognition and recall distinction is the fact that the testing effect (i.e., improved retention from taking a test) in the second study was limited to meaning recognition knowledge but not meaning recall. This was explained in terms of transfer-appropriate processing theory (Morris et al., 1977), which suggests that memory performance is enhanced when the cognitive processes used during learning match those required during testing, as the weekly quizzes were administered in a meaning recognition format. Another factor might be that recall knowledge mastery is more difficult than recognition mastery. Since learning to the recall level likely takes more time than learning to the recognition level, the low study time might additionally explain why gains were limited to meaning recognition. Study time analysis showed that students did not spend much time learning from the app, averaging 0.60 days a week which is nearly three times

less than the average learning time in a similar study (Seibert Hanson & Brown, 2019). These factors could explain the limited gains in meaning recall knowledge.

Overall, the findings of Ozturk (2012) Laufer & Goldstein, (2004), González-Fernández and Schmitt (2019) and the present study show that recall vocabulary knowledge lags behind recognition. The present study is one of the few studies that show this for form-meaning link knowledge using longitudinal data.

6.1.2 Individual differences in vocabulary learning

As might be expected, learners in both studies differed in their learning with some performing very well while others made virtually no progress. Four individual differences have been investigated in this study: out-of-class exposure, motivation, self-regulation and perceptions. The two common across the two studies were motivation and self-regulation.

In terms of out-of-class exposure, Pellicer-Sánchez (2019) notes that previous studies on vocabulary growth (e.g., Webb & Chang, 2012) examined only the effect of in-class exposure on vocabulary learning and did not take into account out-of-class exposure. Study one is the first study that takes into account out-of-class exposure and other key individual differences, namely motivation and self-regulation, to jointly explain variation in vocabulary learning among EFL students using longitudinal data. Several studies have generally shown that out-of-class exposure can lead to vocabulary learning (Arndt & Woore, 2018; Feng & Webb, 2020; Peters, 2018, 2019; Peters & Webb, 2018; see section 3.3.1). Study one investigated seven common sources of out-of-class language exposure (songs, movies, TV programs, games, books, magazines, and websites) using the ESLC questionnaire (Peters, 2018). The two sources that had significant positive effects on meaning recognition vocabulary learning in the Saudi context were playing video games and

listening to songs. Listening to songs in particular had a significant effect on both meaning recognition and meaning recall vocabulary knowledge. Previous experimental research such as Pavia et al. (2019) found that listening to songs can lead to meaning recognition and meaning recall vocabulary learning.

The finding in this study on the positive effects of gaming is also echoed in de Wilde and Eyckman's study (2017) which showed a small but significant positive effect for the amount of gaming on vocabulary test scores. Both listening to songs and playing video games offer additional sources of input for incidental vocabulary learning. Listening to songs can be useful for vocabulary learning given that people tend to listen to the same song multiple times (Abbott, 2002; Conrad et al., 2019) which enhances one of the key conditions of word learning, i.e., repetition (Uchihara et al., 2019; Webb, 2007b). Playing video games also has certain features that help in vocabulary learning, one key of which is the opportunity to develop both receptive and productive vocabulary. Most types of out-of-class exposure (reading, viewing and listening to songs) are receptive in nature and lack interaction which might reduce the opportunities for productive language skills to develop. Video games on the other hand (especially when there is high interaction between the players such as MMORPGs) offer opportunities for productive vocabulary (Janebi Enayat & Haghighatpasand, 2019) and language skills to develop (Jabbari & Eslami, 2023). For example, Jabbari and Eslami (2023) analyzed interactions and negotiation of meaning[12] episodes between MMORPGs players over six months (59.96 hours of recorded audio and nine hours of screen-recorded gaming sessions). They found that playing MMORPGs games offers ample opportunities for comprehensible input, producing comprehensible output and attention to L2 form, all of which

[12] Negotiation of meaning is the process of resolving communication breakdown. SLA research suggests negotiation of meaning brings learners' attention to L2 form which can facilitate language development (Long, 1983).

SLA research suggests as being conducive for language learning (Gass & Mackey, 2007; Gass & Selinker, 2008).

The analysis of students' out-of-class language exposure also revealed which English language input they actually engage with. The activities Saudi EFL students engage with most frequently are (from most to least frequent) playing video games, watching movies with L1 subtitles, listening to songs and watching series with L1 subtitles. Despite cultural differences, the type and frequency of Saudi EFL out-of-class exposure are very similar to Flemish EFL learners (Peters, 2018). For example, the most frequent activities in both groups included watching movies and shows, listening to songs and playing video games, while the least frequent activities were reading books and magazines. This might be one impact of globalization in which the widespread availability of media and technology have contributed to a global culture (Crystal, 2017). Movies and video games for instance are often produced and distributed globally reaching diverse populations (Godwin-Jones, 2018). This exposure seems to create shared experiences and interests among teenagers across different cultures. Additionally, the fact that young Saudi and Flemish EFL learners prefer spending time online (e.g., watching movies and playing games) more than reading books seems to suggest that they find these activities more fun and entertaining. The activities that are both effective and enjoyable are more likely to be useful for vocabulary learning than the activities that are only effective (e.g., reading in this study) primarily due to motivation. When learners derive pleasure from an activity, they are more inclined to invest time in it. This aligns with Krashen's optimal input hypothesis (1982), which argues that comprehensible input alone is not enough for language learning and that input should also be interesting. Meanwhile, the findings of the current study do not align with his view (Krashen, 2004) that language learners and young people in general find reading for pleasure enjoyable as evident by their reported low reading rates.

The decline of reading among younger people is also present in other countries such as the US and the UK. In the US, there was nearly a 50% decrease in reading among children in 2020 compared to 1984 which is the least in decades (National Center for Education Statistics, 2020). Reading for pleasure is also at record lowest among children and young people in the UK with a 38% decline in reading for pleasure in 2023 (National Literacy Trust, 2023). Smartphones and social media are among the main cited reasons for this decline.

The fact that extensive reading is not very common among Saudi and Flemish learners brings to attention a potential limitation. Enthusiastic views on the advantages of extensive reading such as Nation's (2022, p. 590) suggestion that "[t]he single most effective improvement that a teacher could make to a course on learning English as a foreign language is to include an extensive reading program" perhaps need to take into account the limitation above. In addition, previous experimental research on extensive reading in the Saudi context did not show remarkable results, as shown in Al-Homoud and Schmitt's study (2009) which showed that extensive reading was not superior to intensive reading. The lack of appeal for L2 extensive reading might be because reading for pleasure in either L1 or L2 is not common among Saudi individuals (Al-Homoud & Schmitt, 2009; AL-Qahtani, 2016; Alroqi et al., 2022; GASTAT, 2018). Overall, the finding that extensive reading was the activity students least engaged with and the findings from Al-Homoud and Schmitt's study (2009) in addition to the low rates of reading for pleasure makes it less certain that extensive reading will be "the single most effective improvement" in the Saudi EFL context and perhaps in other contexts. Extensive viewing appears to be a viable option given that most students in this study reported more engagement with watching movies and TV shows (see section 2.3.1.2 for some points to consider).

Motivation is an important factor that could explain why some students in both studies learned more vocabulary than others. Several studies have shown that motivation is a strong predictor of vocabulary achievement (Alamer, 2021a; Elley, 1989; Gardner et al., 1985; Tseng & Schmitt, 2008) as well as more general language skills (Jodai et al., 2014; Spolsky, 2000; see section 3.3.3). Both studies used the self-determination theory (Ryan & Deci, 2017) to better understand the role of motivation in vocabulary learning. As discussed in section 2.3.3, the theory conceptualizes motivation as a series of orientations along a continuum from non-self-determined to self-determined. It makes a key distinction between autonomous motivation (doing something because it's personally rewarding or important) and controlled motivation (doing something due to personal obligation, external reward or to avoid external punishment).

Autonomous motivation had a significant positive effect on meaning recall vocabulary in both studies. The finding is consistent with research on vocabulary that higher levels of autonomous motivation seem to lead to more vocabulary learning (Alamer, 2018; J. H. Lee et al., 2022; Y. Zhang et al., 2017). At the same time, there were no significant effects for autonomous motivation on meaning recognition vocabulary knowledge in both studies. This could be related to recall vocabulary learning being more difficult than recognition vocabulary learning (González-Fernández & Schmitt, 2019; Laufer & Goldstein, 2004). More autonomously motivated learners seem to be better distinguished from less motivated learners when the task is more difficult (Kyndt et al., 2011). Jurczyk et al., p. (2019, p. 295) points out that "[d]iverse psychological theories suggest a direct link between task difficulty and motivation, namely a dynamic increase in motivation with increasing task difficulty". They reviewed studies from cognitive psychology, motivational and social psychology and the neurosciences showing that motivation varies dynamically with task complexity (Jurczyk et al., 2019). The increased difficulty associated

with meaning recall tests could have been the reason why autonomous motivation showed an effect on meaning recall tests scores but not on meaning recognition tests scores in this study.

Most previous studies found either negative or no effect of controlled motivation on language learning outcomes (Alamer, 2021a; Liu, 2007; Noels et al., 1999; F. X. Wang, 2008). The results of the present study match those studies in that controlled motivation had a significant negative effect on the weekly and biweekly groups in the second study and no significant effects in the first study. The findings of both studies generally lend support to the self-determination theory conceptualization of motivation (Ryan & Deci, 2017). Students in both studies who learn English because it is intrinsically rewarding or desirable managed to learn vocabulary significantly more to the meaning recall level. On the other hand, students in the second study who reported learning English for external reasons such as guilt or grades learned significantly less vocabulary on both meaning recognition and meaning recall levels.

The comprehensive models in both studies did not show significant effects for self-regulation on vocabulary learning as measured by the SRCvoc instrument (Tseng et al., 2006). One possible reason for this could be the instrument used here. Several studies have reported issues establishing the construct validity of the SRCvoc with making significant changes to the factors structure (Mizumoto & Takeuchi, 2011; Yeşilbursa & Bilican, 2013). For example, Mizumoto and Takeuchi (2011) obtained poor model fit for the SRCvoc when was used with 443 EFL learners in Japan. They only achieved good model fit after removing two of the five instrument constructs (i.e., commitment control and satiation control). The fact that several studies failed to establish the construct validity of the SRCvoc instrument without major modifications indicates potentially to issues in the instrument. The fact that both study 1 and study 2 showed no effects for SRCvoc could be due to the use of SRCvoc as a measure of self-regulation.

Overall, the findings on the role of individual differences in vocabulary learning from the two thesis studies showed that students who learned more vocabulary throughout the semesters tended generally to have higher out-of-class exposure to English (as study one showed) and higher motivation levels (as found in both studies). Self-regulation, on the other hand, did not predict vocabulary learning in both the first and the second study (in the comprehensive models) possibly due to instrument adaptation (see study 1 discussion, section 4.6).

6.1.3 Textbooks

The first study showed that secondary students had significantly less vocabulary growth compared to intermediate students despite having one more year of instruction and having an extra class every week. One potential factor that has been discussed in this thesis is the possibility that the school textbooks might be hindering the vocabulary growth of secondary students by not introducing enough new words at this stage. Alsaif and Milton (2012, p. 28) suggest that:

> "The decision by the textbook writers to diminish the volumes of vocabulary input after this point [intermediate years] appears very short-sighted. If learners are not expanding their vocabulary at this stage [secondary years], then they may well not progress in their language learning overall".

Given that textbooks are likely the major source of vocabulary learning in an EFL classroom (due to limited input; Jordan & Gray, 2019; Milton & Vassiliu, 2000), it is not improbable that they could hinder vocabulary development. The implications of this finding and other findings are discussed in the next section.

6.2 Implications

The thesis offers important implications for vocabulary pedagogy and research. The main implication of study one is that vocabulary growth in an EFL context mostly does not take care of itself, as assumed by some language teachers, textbook developers and researchers (Bergström et al., 2021; Laufer, 2003, 2005, 2006, 2009; Nation, 2022; Schmitt & Schmitt, 2020; Webb & Nation, 2017) and that even after many years of school instruction, students might still not develop a good knowledge of the highest frequency vocabulary (i.e., the most frequent 1000 word-families). Vocabulary, as suggested by vast research (Clenton & Booth, 2020; Nation, 2022; Schmitt, 2020; Schmitt et al., 2017; see section 2.1.4) is central for language comprehension and production and a lack of substantial vocabulary progress is a serious indicator of a lack of overall language progress (Milton, 2009; Nation, 2022; Schmitt & Schmitt, 2020).

To address this low development in high frequency vocabulary, an intentional vocabulary learning component should be included in an EFL language learning program to assist learners in mastering the highest frequency words (Webb & Chang, 2012; Webb & Nation, 2017). This can take several forms, the most effective of which for retention are flashcards and wordlists (see section 6.1.1; Webb et al., 2020). Flashcard learning can be implemented in-class, but if classroom time is limited, which is the case in many EFL contexts (Lightbown & Spada, 2020), it can be assigned as an out-of-class activity. One of the main implications of the second study is that students might not engage with out-of-class language learning even if course credits were used. Therefore, it recommended, based on the findings of the second study, to supplement out-of-class flashcard learning with in-class quizzes as this led to significant vocabulary learning.

The second main implication of study two is the finding that more frequent quizzes do not seem to lead to significantly more vocabulary learning. Based on this finding, it seems that teachers can

choose lower frequency quizzes without missing out on significant vocabulary gains. More frequent quizzes add extra workload to the already busy schedule of many language teachers. Therefore, from a strictly vocabulary gain perspective, it might be more practical to give students fewer quizzes as they appear to result in similar gains to more frequent quizzes. Although no additional vocabulary gains were found for more frequent quizzes, there are some advantages associated with more frequent quizzes that might make them desirable such as increased class attendance (Schrank, 2016), which has a relatively strong and positive association with class grades (Credé et al., 2010). On the other hand, more frequent quizzes do have some issues such as taking time away from learning (Roediger et al., 2011) which might be a problem given the limited time available for foreign language instruction in many contexts. Instructors should consider these and other factors when deciding on the optimal quiz frequency that suits their context. Apart from these main implications, there are other important implications worth discussing. These focus on out-of-class exposure, motivation, promoting recall mastery of words and improving how vocabulary is treated in textbooks.

6.2.1 Out-of-class exposure and motivation

Students should be encouraged to increase their exposure to English out-of-class through activities they prefer such as playing video games, watching movies and listening to songs, which have the potential to promote their incidental vocabulary learning (Peters, 2018, 2019; Peters & Webb, 2018). Seeking and engaging with opportunities for vocabulary learning in out-of-class settings requires motivated students. Both studies highlighted the importance of autonomous motivation in vocabulary learning. For learners to engage more effectively in intentional and incidental vocabulary learning, they need to be autonomously motivated. Students' autonomous motivation needs to be nourished by fulfilling their three basic psychological needs of autonomy, competence

and relatedness (Alamer, 2021a; Noels et al., 2019). Jones et al. (2009) suggest a number of activities (e.g., creating assignments of varying difficulties) that can be applied in foreign language classrooms to promote learners' autonomous motivation. Within the context of vocabulary learning from digital flashcards, students' need for autonomy can be nourished by, for example, allowing them to choose which words they wish to learn to make their learning more personalized (while making sure they are aware of the importance of high and mid-frequency words). Students' sense of competence could be developed by starting with a small number of words to learn at the beginning and then they can work their way up by making incremental increases. Finally, students' sense of relatedness could be fostered by, for example, allowing peer assessment (Wilkinson, 2020) in which students grade the quizzes of their classmates, leading to more collaborative learning.

6.2.2 Promoting recall mastery

Recall mastery of words is important for language use (Nation, 2022) and might be a better predictor of reading comprehension than word recognition (McLean et al., 2020; Stewart et al., 2024; Stoeckel et al., 2019). Since the results of the first study showed that meaning recall vocabulary knowledge growth lags behind meaning recognition, one effective way of enhancing recall memory of words may be to use retrieval during learning (Barcroft, 2007; Nakata, 2016) as in flashcard learning (see section 5.1.1). The second study showed significant vocabulary learning to the meaning recall level when the quiz groups were combined together and compared to the no-quiz group. However, the gains were not very robust (they were nonsignificant when each group was analyzed individually), potentially due to the fact that students in general did not spend much time learning from the app. The time students spent learning words from the app might not have been enough to generate deeper learning that is required for recall retrieval. In addition to

flashcards, studies suggest other techniques to enhance recall such as the use of vocabulary strategy instruction (e.g., linking a new word to a previously learned word) which might help in learning vocabulary to the recall level (Atay & Ozbulgan, 2007).

In terms of theory, any conceptualization of vocabulary knowledge development needs to take into account the findings from both studies that meaning recognition and meaning recall knowledge develop differently (González-Fernández & Schmitt, 2019). Meaning recall growth was three times smaller than meaning recognition growth in study one and two times smaller in study two. These findings need to be considered when developing a model or a theory of how vocabulary knowledge develops in L2.

6.2.3 Textbooks

Secondary textbooks in Saudi Arabia need to be reviewed to check if they provide adequate vocabulary because the findings of study 1 showed that the secondary students made very little vocabulary growth compared to the intermediate students. This warrants a review because the same issue of introducing little vocabulary noted by Alsaif and Milton (2012) a while ago seems to persist today. Milton and Hopwood (2022) note that many language learning textbooks lack a principled approach to vocabulary learning, which could be the case in the Saudi context. A similar situation appears to exist in the UK context but on the level of the curriculum. Milton and Hopwood (2022) showed that relatively little vocabulary is being introduced in the UK foreign language curriculum (1200-1700 words), based on the idea that focusing on high frequency words only is enough to communicate in a wide variety of situations which will save teaching time and make language learning more accessible. Milton and Hopwood (2022) rightly argue that the choice of minimizing vocabulary to a small set misses the fact that vocabulary in the thousands is needed for language learners to use language receptively and productively (see section 2.1.4). The very

limited vocabulary development of secondary students (study 1) needs attention from the Saudi Ministry of Education. This is to understand more about why it is happening in the sample school across two years and whether the same situation exists in other secondary schools.

Overall, both thesis studies provide novel and important implications for pedagogy and research. Three are of major importance. First, EFL students after many years might not even master the most frequent 1000 word-families. Second, out-of-class flashcard learning supplemented with in-class quizzes can be an effective approach to help students struggling with learning high frequency vocabulary. Finally, it seems that teachers have the option to conduct quizzes less frequently without a significant decrease in vocabulary gains given that more frequent quizzes did not lead to more vocabulary learning.

6.3 Limitations and future research

The two thesis studies provide valuable findings about how vocabulary develops in a foreign language context and how this development can be enhanced. The findings nonetheless need to be interpreted with some limitations in mind. Suggestions for future research are embedded in the discussion to help overcome some of these limitations and explore new areas.

The first study explored receptive vocabulary growth (meaning recognition and meaning recall) but did not investigate productive vocabulary growth. Given that the two usually develop differently (Laufer, 1998), future research should examine how different the development of receptive and productive vocabulary is (Pellicer-Sánchez, 2019; Schmitt, 2019). Future research should also go beyond form-meaning knowledge and explore longitudinally how other aspects of vocabulary knowledge (e.g., multiple meanings and collocations) develop over time (González-Fernández & Schmitt, 2020; Schmitt, 2019). Due to time constraints, the bilingual vocabulary tests

were not piloted. Relatedly, the scoring of the meaning recall tests was done by the author only. Future research should include another trained rater during scoring meaning recall tests and calculate inter-rater agreement (e.g., Kappa) to enhance scoring reliability. Due to gender segregation in public education in Saudi Arabia, the participants were male students only. The limited research on the effect of gender on vocabulary knowledge development shows mixed findings with some studies showing significant differences (Alqarni, 2019) while others showing nonsignificant differences (S. Lee, 2020; Simos et al., 2012). For instance, Lee (2020) found no significant gender differences in vocabulary breadth and depth among Korean EFL learners. In contrast, Alqarni (2019) found that Saudi male students scored significantly higher on all levels of the vocabulary levels test. In terms of the individual differences, female students might be different in aspects such as their preference for out-of-class activities. Girls for instance spend often less time playing video games compared to boys (Sundqvist & Wikström, 2015). Future work should aim to collect data from both male and female schools to enhance the generalizability of findings. Similarly, although Saudi schools across the kingdom are mostly similar (e.g., similar hours of English instruction; Al-Hoorie, Shlowiy, et al., 2021) and the vocabulary sizes of secondary students found in study 1 were very similar to previous studies conducted in other cities (see section 4.6), other factors such as higher socio-economic status (SES) in other major Saudi cities (e.g., Jeddah or Dammam) might lead to higher EFL proficiency (Huang et al., 2018). More data from other Saudi cities will be useful to check the generalizability of the first study findings and reduce uncertainty.

The second study has shown that digital flashcard learning can be significantly improved when paired with in-class quizzes. Similar to study one, it investigated the form-meaning link knowledge only. Although it is the most important aspect of vocabulary knowledge (Laufer & Goldstein,

2004), there are other important aspects of knowing a word such as knowledge of the spoken form, multiple meanings and derivations. For knowledge of spoken form for instance, future research can focus on making quizzed flashcard learning more effective by for example using spoken quizzes instead of written quizzes. This might help in learning both the spoken and written form of words. A study that compared the effect of the two types of quizzes (spoken vs. written) on vocabulary learning from a wordlist found that spoken quizzes led to significantly higher scores on a final vocabulary listening test (Uchihara, 2023). At the same time, the scores of both groups were similar on a final written test suggesting similar gains. The second study, like the first, focused on receptive knowledge and did not include productive knowledge measures. Future research should explore whether the receptive vocabulary gains found in this study extend to productive knowledge and other aspects of vocabulary knowledge.

Being conducted in classroom settings, there was a three-week break during the experiment (see Figure 14) which might have some effect on the study. Future work should aim to run the experiment over consecutive weeks if possible to minimize any influence on learning.

The three quiz groups varied in terms of the number of items tested. The weekly group was quizzed on all 120 items, the biweekly on 60 items and the monthly on 30 items. It is not possible to control for the number of quizzed items while at the same time control for the number of items appearing on the quizzes. Future research should examine the effect of quizzing all groups on all items while allowing the length of the frequent quizzes to vary (e.g., number of items on weekly quizzes would be 15, biweekly would be 30 items and the monthly quizzes would consist of 60 items). The downside of this design is that the less frequent quizzes would take much longer time to complete given that they consist of more items (i.e., weekly = 15 and monthly = 60 items).

Due to technological limitations, the data on how much time students spent learning per week was not very accurate. This information would have provided useful insights as it would have contributed to answering questions like, do students in the biweekly and monthly groups space their learning? Does study time increase or decrease every week? Future research could use an alternative app such as Anki since it provides these kinds of information while keeping in mind its limitations (see section 5.2.3).

The comprehensive models in both studies did not show significant effects for self-regulation on vocabulary learning as measured by the SRCvoc instrument (Tseng et al., 2006). One possible reason for this could be due to the instrument used here and the lack of adaption in the Saudi context in this thesis. Several studies have reported issues establishing the construct validity of the SRCvoc without making significant changes to the factors structure (Mizumoto & Takeuchi, 2011; Yeşilbursa & Bilican, 2013). For example, Mizumoto and Takeuchi (2011) obtained poor model fit for the SRCvoc when was used with 443 EFL learners in Japan. They only achieved good model fit after removing two of the five instrument constructs (i.e., commitment control and satiation control) which is considered major changes that make it less clear whether the findings provide support for the construct validity of SRCvoc (Alamer et al., 2024). The fact that several studies failed to establish the construct validity of the SRCvoc instrument without major modifications indicates potentially to issues in the instrument and the need for extensive adaptation. Others however argue that the issue with SRCvoc adaptation is methodological and argue for the use of a different validation approach such as confirmatory composite analysis instead of confirmatory factor analysis (Alamer et al., 2024). Alamer et al. (2024) obtained good model fit for SRCvoc though confirmatory composite analysis but not confirmatory factor analysis without modifications to the instrument factors structure. Nevertheless, the lack of significant effects for

SRCvoc in both study 1 and 2 and the findings from other adaption studies indicate that need for future research to conduct extensive adaption before using the instrument in a new context.

While the second study considered key factors such as individual differences, there are other factors that were not covered that may moderate the effect of quizzes on digital vocabulary learning. These can include for instance the percentage of class credit allocated for digital flashcard learning and the number of words to learn. A useful starting point perhaps is to examine the moderators that were found significant in meta-analyses on the testing effect (Yang et al., 2021). Some of the significant moderating factors found include whether feedback was offered and whether a test was administered in or out-of-class. Knowing more about these factors might be useful in raising awareness of potential confounds and perhaps in mitigating their negative effects. It can also highlight some potential learning enhancers (such as giving feedback) which might make learning more effective.

6.4 Conclusion

The thesis has provided a number of findings and recommendations that have the potential to improve L2 vocabulary learning in EFL contexts. Study 1 has shown that vocabulary learning can be very limited in foreign language contexts and leaners may not even master the most frequent 1000 words after many years of instruction. Those who thrive in this context tend to be more autonomously motivated and have higher out-of-class exposure to English. Study 2 has shown that vocabulary learning can be significantly enhanced by supplementing flashcard learning with in-class quizzes. It has also shown that more frequent vocabulary quizzes do not necessarily lead to more vocabulary learning.

What is important in the next stage is for these findings to find their way into the classroom. This concern is warranted since there appears to be a gap between SLA researchers and teachers (Borg, 2010; R. Ellis, 2010; Sato & Loewen, 2019a, 2022; Spada, 2015; Spada & Lightbown, 2022). This gap can lead teachers to depend on their personal experiences which might not lead to optimal language learning (Sato & Loewen, 2022). One reason for this gap is the perceived lack of relevance of some SLA research to the classroom (Spada, 2015). SLA researchers are described as being "removed from day-to-day classroom practice and oriented to more abstract variables" (Spada & Lightbown, 2022, p. 635). This position seems to be reinforced by the finding that only one third of research published in two SLA journals (Studies in Second Language Acquisition and Language Learning) from 1990 to 2010 was conducted in regular classroom contexts (Plonsky, 2013). The present study aimed to overcome this by conducting both studies in naturalistic classroom settings to increase the ecological validity of the findings while adhering to scientific rigor as much as possible (Sato & Loewen, 2019b).

Another reason for the gap between research and practice is due to teachers' limited access to SLA research. The three main barriers are the lack of physical access to research (e.g., expensive paywalls to read journal articles), the low readability of articles due to the technical nature of academic writing and teachers' lack of time to read research (Sato & Loewen, 2019a). One solution that has been proposed to tackle these obstacles is the Open Accessible Summaries in Language Studies (OASIS) initiative (Alferink & Marsden, 2023). OASIS aims to bridge the gap between research and pedagogy by allowing researchers to share a non-technical summary of their research which covers the topic of study, its importance, the methodology used and the findings. OASIS has more than 1350 articles with 8-12 articles added every week (Alferink & Marsden, 2023). The fact that these summaries are free, non-technical and short might help minimize the research-

practice gap. In terms of the current thesis, a non-technical and plain summary of the findings made in this thesis will be shared on the OASIS website and on social media platforms to maximize teachers' accessibility.

In conclusion, the thesis overall makes valuable contributions to both vocabulary theory and practice. The first study enhances our understanding of the nature of vocabulary knowledge by examining vocabulary growth longitudinally. The second study offers practical recommendations to help language learners learn vocabulary more effectively. The two studies combined make important strides in advancing L2 vocabulary research.

REFERENCES

Abbott, M. (2002). Using music to promote L2 learning among adult learners. *TESOL Journal*, *11*(1), 10–17. https://doi.org/10.1002/j.1949-3533.2002.tb00061.x

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Aghlara, L., & Tamjid, N. H. (2011). The effect of digital games on Iranian children's vocabulary retention in foreign language acquisition. *Procedia - Social and Behavioral Sciences*, *29*, 552–560. https://doi.org/10.1016/j.sbspro.2011.11.275

Agustín-Llach, M. P., & Alonso, A. C. (2016). Vocabulary growth in young CLIL and traditional EFL learners: Evidence from research and implications for education. *International Journal of Applied Linguistics*, *26*(2), 211–227. https://doi.org/10.1111/ijal.12090

Ahmed, M. O. (1989). Vocabulary learning strategies. In P. Meara (Ed.), *Beyond words* (pp. 3–14). CILT.

Al fotais, A. (2019). *Investigating the effect of spaced versus massed practice on vocabulary retention in the EFL classroom [Doctoral dissertation]* [University of Essex]. http://repository.essex.ac.uk/id/eprint/25062

Alamer, A. (2018). *Saudi students' English achievement as a second language: A motivational process model linking basic psychological needs, self-determination theory, goal orientation, and motivational emotion [Doctoral dissertation]*. University of Sydney.

Alamer, A. (2021a). Basic psychological needs, motivational orientations, effort, and vocabulary knowledge: A comprehensive model. *Studies in Second Language Acquisition*, 1–21. https://doi.org/10.1017/S027226312100005X

Alamer, A. (2021b). Construct validation of self-determination theory in second language scale: The bifactor exploratory structural equation modeling approach. *Frontiers in Psychology*, *12*(September), 1–7. https://doi.org/10.3389/fpsyg.2021.732016

Alamer, A., Teng, M. F., & Mizumoto, A. (2024). Revisiting the construct validity of self-regulating capacity in vocabulary learning scale: The confirmatory composite analysis (CCA) approach. *Applied Linguistics*, 1–18.

Alfairouz, A. (2015). *Measuring receptive and productive vocabulary sizes of EFL English teachers in public schools in Saudi Arabia [Doctoral dissertation]*. Swansea University.

Alferink, I., & Marsden, E. (2023). OASIS: One resource to widen the reach of research in language studies. *Innovation in Language Learning and Teaching*, *17*(5), 946–952. https://doi.org/10.1080/17501229.2023.2204100

Alhaj, A. A. M., Alwadai, M. A. M., & Albuhairi, H. M. (2019). Evaluating Saudi EFL secondary schools students' performance on Paul Nation's standardized vocabulary level tests. *LLT Journal: A Journal on Language and Language Teaching*, *22*(1), 126–136.

Al-Hazemi, H. A. A. G. (1993). *Low-level EFL vocabulary tests for Arabic speakers [Doctoral dissertation]*. Swansea University.

Al-Homoud, F., & Schmitt, N. (2009). Extensive reading in a challenging environment: a comparison of extensive and intensive reading approaches in Saudi Arabia. *Language Teaching Research*, *13*(4), 383–401. https://doi.org/10.1177/1362168809341508

Al-Hoorie, A. H., Al-Shahrani, M., Al-Shlowiy, A., & Mitchell, C. (2021). The preparation of teachers of EAL in Saudi Arabia: Research, policy, curriculum and practice. In N. Polat, L. Mahalingappa, & H. Kayi-Aydar (Eds.), *The preparation of teachers of English as an additional language around the world* (pp. 158–187). Multilingual Matters. https://doi.org/10.21832/9781788926164-010

Al-Hoorie, A. H., Shlowiy, A. Al, & Mitchell, C. (2021). Preparation of teachers of EAL in Saudi Arabia:Research, policy, curriculum and practice. In N. Polat, L. Mahalingappa, & H. Kayi-Aydar (Eds.), *The preparation of teachers of English as an additional language around the world: Research, policy, curriculum and practice* (Issue January, pp. 158–187). https://doi.org/10.21832/polat6157

Ali, B. M. (2023, June 30). How many years does a typical user spend on social media? *Al Jazeera News*, 1–10. https://www.aljazeera.com/news/2023/6/30/how-many-years-does-a-typical-user-spend-on-social-media

Aljumah, F. (2011). Developing Saudi EFL students' oral skills: An integrative approach. *English Language Teaching*, *4*(3), 84–89. https://doi.org/10.5539/elt.v4n3p84

Al-Khasawneh, F. (2019). The impact of vocabulary knowledge on the reading comprehension of Saudi EFL learners. *Journal of Language and Education*, *5*(3), 24–34. https://doi.org/10.17323/jle.2019.8822

Allen, D. (2018). Cognate frequency and assessment of second language lexical knowledge. *International Journal of Bilingualism*, *23*(5), 1121–1136. https://doi.org/10.1177/1367006918781063

Allen, D. (2019a). Cognate frequency predicts accuracy in tests of lexical knowledge. *Language Assessment Quarterly*, *16*(3), 312–327. https://doi.org/10.1080/15434303.2019.1635134

Allen, D. (2019b). The prevalence and frequency of Japanese-English cognates: Recommendations for future research in applied linguistics. *International Review of Applied Linguistics in Language Teaching*, *57*(3), 355–376. https://doi.org/10.1515/iral-2017-0028

Allen, D. (2020). An overview and synthesis of research on English loanwords in japanese. *Vocabulary Learning and Instruction*, *9*(1), 33–50. https://doi.org/10.1177/1367006918781063

Almansour, A. (2019). *An investigation into the use of the smartphone application "Memrise" in supporting English vocabulary learning among undergraduate students in Saudi Arabia [Doctoral dissertation]*. Liverpool John Moores University.

Al-Masrai, A., & Milton, J. (2012). The vocabulary knowledge of university students in Saudi Arabia. *TESOL Arabia Perspectives*, *19*(3), 13–19. www.tesolarabia.org

Almutairi, N. (2008). *The influence of educational and sociocultural factors on the learning styles and strategies of female students in Saudi Arabia [Doctoral dissertation]*. University of Leicester.

Al-Nujaidi, A. H. (2003). *The relationship between vocabulary size, reading strategies, and reading comprehension of EFL learners in Saudi Arabia [Doctoral dissertation]*. Oklahoma State University.

AL-Qahtani, A. A. (2016). Why do Saudi EFL readers exhibit poor reading abilities? *English Language and Literature Studies*, *6*(1), 1. https://doi.org/10.5539/ells.v6n1p1

Alqarni, I. R. (2018). Saudi English major freshmen students' vocabulary learning strategies: An exploratory study. *International Journal of Applied Linguistics and English Literature*, *7*(1), 141–145. https://doi.org/10.7575/aiac.ijalel.v.7n.1p.141

Alqarni, I. R. (2019). Receptive vocabulary size of male and female Saudi English major graduates. *International Journal of English Linguistics*, *9*(1), 111. https://doi.org/10.5539/ijel.v9n1p111

Alqurashi, N. (2020). *The relationship between vocabulary size and training in vocabulary-learning strategies: A case study of preparatory year students at Saudi university [Doctoral dissertation]*. University of Leicester.

Alrabai, F. (2016). Factors underlying low achievement of Saudi EFL learners. *International Journal of English Linguistics*, *6*(3), 21. https://doi.org/10.5539/ijel.v6n3p21

Alrabai, F. (2017). Exploring the unknown: The autonomy of Saudi EFL learners. *English Language Teaching*, *10*(5), 222. https://doi.org/10.5539/elt.v10n5p222

Alrabai, F. (2018). Learning English in Saudi Arabia. In C. Moskovsky & M. Picard (Eds.), *English as a foreign language in Saudi Arabia: New insights into teaching and learning English*. Routledge.

Al-Rasheed, M. (2010). *A history of Saudi Arabia*. Cambridge University Press.

Alroqi, H., Serratrice, L., & Cameron-Faulkner, T. (2022). The home literacy and media environment of Saudi toddlers. *Journal of Children and Media*, *16*(1), 95–106. https://doi.org/10.1080/17482798.2021.1921819

Alsaedi, A. E. (2012). *The teaching of EFL speaking in developed secondary public schools for females in Saudi Arabia: A case study [Doctoral dissertation]*. University of Southampton.

Alsaif, A. (2011). *Investigating vocabulary input and explaining vocabulary uptake among EFL learners in Saudi Arabia (Doctoral dissertation)*. Swansea University.

Alsaif, A., & Masrai, A. (2019). Extensive reading and incidental vocabulary acquisition: The case of a predominant language classroom input. *International Journal of Education and Literacy Studies*, *7*(2), 39. https://doi.org/10.7575/aiac.ijels.v.7n.2p.39

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *The Language Learning Journal*, *40*(1), 21–33. https://doi.org/10.1080/09571736.2012.658221

Al-Seghayer, K. (2011). *English teaching in Saudi Arabia: Status, issues, and challenges.* Hala.

Alshammari, M. M. (2011). The use of the mother tongue in Saudi EFL classrooms. *Journal of International Education Research (JIER)*, *7*(4), 95–102. https://doi.org/10.19030/jier.v7i4.6055

Alsharif, R. (2022). Relationship between Vocabulary Learning Strategies and Vocabulary Size: Evidence from Saudi Female EFL Learners. *International Journal of Education and Literacy Studies*, *10*(1), 188. https://doi.org/10.7575/aiac.ijels.v.10n.1p.188

Alsuhaibani, Y., Altalhab, S., Borg, S., & Alharbi, R. (2023). 15 years' experience of teaching English in Saudi Primary Schools: Supervisors' and teachers' perspectives. *Linguistics and Education*, *77*(August), 101222. https://doi.org/10.1016/j.linged.2023.101222

Altalhab, S. (2014). *Teaching and learning vocabulary through reading at Saudi universities [Doctoral dissertation]*. University of Strathclyde.

Altalhab, S. (2019). The vocabulary knowledge of Saudi EFL tertiary students. *English Language Teaching*, *12*(5), 55. https://doi.org/10.5539/elt.v12n5p55

Alzahrani, S. A. (2020). Relationship between vocabulary size and reading comprehension achievement among Saudi high school EFL relationship between vocabulary size and reading. *Hamdard Islamicus*, *43*(2), 2281–2288. https://hamdardfoundation.org/hamdard

Amlund, J. T., Kardash, C. A. M., & Kulhavy, R. W. (1986). Repetitive reading and recall of expository text. *Reading Research Quarterly*, *21*(1), 49–58. https://doi.org/10.2307/747959

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). International Reading Association Inc.

Andringa, S., & Godfroid, A. (2019). *SLA for all? Call for participation*. EuroSLA. https://www.eurosla.org/sla-for-all-call-for-participation/

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Arndt, H. L., & Woore, R. (2018). Vocabulary learning from watching YouTube videos and reading blog posts. *Language Learning & Technology*, *22*(1), 124–142. https://doi.org/https://doi.org/10125/44660

Ashcroft, R. J., Cvitkovic, R., & Praver, M. (2018). Digital flashcard L2 vocabulary learning outperforms traditional flashcards at lower proficiency levels: A mixed-methods study of 139 Japanese university students. *The Eurocall Review*, *26*(1), 14–28.

Atay, D., & Ozbulgan, C. (2007). Memory strategy instruction, contextual learning and ESP vocabulary recall. *English for Specific Purposes*, *26*(1), 39–51. https://doi.org/10.1016/j.esp.2006.01.002

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new computer adaptive test of size and strength (CATSS): Development and validation. *Language Assessment Quarterly*, *16*(3), 345–368. https://doi.org/10.1080/15434303.2019.1649409

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Baddeley, A. (1990). *Human memory*. Erlbaum Associates.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Bailey, S. (2011). Teenagers learning languages out of school: what, why and how do they learn? In P. Benson & H. Reinders (Eds.), *Beyond the language classroom* (pp. 119–31). Palgrave Macmillan.

Bamford, J., & Day, R. R. (2004). *Extensive reading activities for teaching language.* Cambridge University Press.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*(2), 89–99. https://doi.org/10.1080/00220671.1991.10702818

Baranowska, K. (2020). Learning most with least effort: subtitles and cognitive load. *ELT Journal*, *74*(2), 105–115. https://doi.org/10.1093/elt/ccz060

Baranowska, K. (2021). Learning most with least effort: Subtitles and cognitive load. *ELT Journal*, *74*(2), 105–115. https://doi.org/10.1093/ELT/CCZ060

Barclay, S., & Schmitt, N. (2019). Current perspectives on vocabulary teaching and learning. In X. Gao (Ed.), *Second Handbook of Teaching English Language Teaching* (pp. 799–817). Springer. https://doi.org/10.1007/978-3-030-02899-2

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Barnawi, O. Z., & Al-Hawsawi, S. (2017). English education policy in Saudi Arabia: English language education policy in the Kingdom of Saudi Arabia: Current trends, issues and challenges. In R. Kirkpatrick (Ed.), *English language education policy in Middle East and North Africa* (Vol. 13). Springer. https://doi.org/10.1007/978-3-319-46778-8

Barrot, J. S. (2022). Social media as a language learning environment: A systematic review of the literature (2008-2019). *Computer Assisted Language Learning*, *35*(9), 2534–2562. https://doi.org/10.1080/09588221.2021.1883673

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beare, K. (2019). *How many people learn English*. ThoughtCo. https://www.thoughtco.com/how-many-people-learn-english-globally-1210367

Beaulieu, R. P., & Zar, M. C. (1986). The effects of examination frequency on student performance. *Journal of Instructional Psychology*, *13*(2), 81. https://www.proquest.com/scholarly-journals/effects-examination-frequency-on-student/docview/1416364737/se-2?accountid=8630

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Benson, P. (2011). Language learning and teaching beyond the classroom: An introduction to the field. In P. Benson & H. Reinders (Eds.), *Beyond the language classroom* (pp. 7–17). Palgrave Macmillan.

Benson, P. (2015). Commenting to learn: Evidence of language and intercultural learning in comments on youtube videos. *Language Learning and Technology*, *19*(3), 88–105.

Bergström, D., Norberg, C., & Nordlund, M. (2021). "Words are picked up along the way" – Swedish EFL teachers' conceptualizations of vocabulary knowledge and learning. *Language Awareness*, *31*(4), 393–409. https://doi.org/10.1080/09658416.2021.1893326

Bisson, M. J., Van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, *35*(2), 399–418. https://doi.org/10.1017/S0142716412000434

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, *14*(1), 14–31. https://doi.org/10.1080/10904018.2000.10499033

Boo, Z., Dörnyei, Z., & Ryan, S. (2015). L2 motivation research 2005–2014: Understanding a publication surge and a changing landscape. *System*, *55*, 145–157. https://doi.org/10.1016/j.system.2015.10.006

Borg, S. (2010). Language teacher research engagement. In *Language Teaching* (Vol. 43, Issue 4). https://doi.org/10.1017/S0261444810000170

Botes, E., Dewaele, J. M., & Greiff, S. (2022). Taking stock: A meta-analysis of the effects of foreign language enjoyment. *Studies in Second Language Learning and Teaching*, *12*(2), 205–232. https://doi.org/10.14746/ssllt.2022.12.2.3

Brevik, L. M. (2019). Gamers, surfers, social Media users: Unpacking the role of interest in English. *Journal of Computer Assisted Learning*, *35*(5), 595–606. https://doi.org/10.1111/jcal.12362

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2022). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, *43*(3), 596–602. https://doi.org/10.1093/applin/amaa061

Brown, J. (1976). *Recall and recognition*. John Wiley & Sons.

Brown, R., Waring, R., & Sangrawee Donkaewbua, J. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*(2), 136–163. http://nflrc.hawaii.edu/rfl

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*(JUL), 1–11. https://doi.org/10.3389/fpsyg.2016.01116

Center for Self-Determination Theory. (2024). *Self-Determination Theory*. https://selfdeterminationtheory.org

Chen, C., & Liu, Y. (2020). The role of vocabulary breadth and depth in IELTS academic reading tests. *Reading in a Foreign Language*, *32*(1), 1–27.

Chen, H.-J. H., & Hsu, H.-L. (2019). The impact of a serious game on vocabulary and content learning. *Computer Assisted Language Learning*, *33*(7), 811–832. https://doi.org/10.1080/09588221.2019.1593197

Clark, M. K., & Ishida, S. (2005). Vocabulary knowledge differences between placed and promoted EAP students. *Journal of English for Academic Purposes*, *4*(3), 225–238. https://doi.org/10.1016/j.jeap.2004.10.002

Clenton, J., & Booth, P. (2020). *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (J. Clenton & P. Booth, Eds.). Routledge. https://doi.org/10.4324/9780429285400

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *The Canadian Modern Language Review*, *57*(2), 295–324. https://doi.org/10.3138/cmlr.57.2.295

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning and Technology*, *11*(3), 38–63.

Cobb, T., & Horst, M. (2011). Does word coach coach words? *CALICO Journal*, *28*(3), 639–661. https://doi.org/10.11139/cj.28.3.639-661

Cohen, A. (1996). *Second language learning and use strategies: Clarifying the issues* (Issue July). University of Minnesota.

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Conrad, F., Corey, J., Goldstein, S., Ostrow, J., & Sadowsky, M. (2019). Extreme re-listening: Songs people love...and continue to love. *Psychology of Music*, *47*(2), 158–172. https://doi.org/10.1177/0305735617751050

Cox, J. G. (2019). Multilingualism in older age: A research agenda from the cognitive perspective. *Language Teaching*, *52*(3), 360–373. https://doi.org/10.1017/S0261444819000193

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A., & Boutorwick, T. J. (2018). Longitudinal vocabulary development in an EMI international school context: learners and texts in EAL, maths, and science. *TESOL Quarterly*, *52*(3), 588–610. https://doi.org/10.1002/tesq.450

Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. In *Review of Educational Research*. https://doi.org/10.3102/0034654310362998

Cronbach, L. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow.* Lawrence Erlbaum Associates Inc.

Crystal, D. (2017). *English as a global language*. Cambridge University Press.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*(3), 369–382. https://doi.org/10.1177/0267658312443651

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and Assessing Vocabulary Knowledge*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667268

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta*, *49*(1), 67–77. https://doi.org/10.7202/009021ar

Dang, T. N. Y. (2021). Selecting lexical units in wordlists for EFL learners. *Studies in Second Language Acquisition*, *43*(5), 954–957. https://doi.org/10.1017/S0272263121000681

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In P. Nation (Ed.), *Making and Using Word Lists for Language Learning and Testing* (pp. 153–167). John Benjamins Publishing Company. https://doi.org/10.1075/z.208.15ch15

Dang, T. N. Y., & Webb, S. (2020). Vocabulary and good language teachers. *Lessons from Good Language Teachers*, *May*, 203–218. https://doi.org/10.1017/9781108774390.019

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords*. Multilingual Matters.

Davie, N., & Hilber, T. (2015). Mobile-assisted language learning: Student attitudes to using smartphones to learn English vocabulary. *Proceedings of the 11th International Conference on Mobile Learning*, 70–78.

Day, R. R., & Robb, T. (2015). Extensive reading. In D. Nunan & J. Richards (Eds.), *Language learning beyond the classroom*. Routledge.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2019). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition*, *23*(1), 171–185. https://doi.org/10.1017/S1366728918001062

De Wilde, V., & Eyckmans, J. (2017). Game on! Young learners' incidental language learning of English prior to instruction. *Studies in Second Language Learning and Teaching*, *7*(4), 673–694. https://doi.org/10.14746/ssllt.2017.7.4.6

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109–134. https://doi.org/10.1016/0092-6566(85)90023-6

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Deck, W. (1998). *The effects of frequency of testing on college students in a principles of marketing course [Doctoral dissertation]*. Virginia Polytechnic Institute and State University.

Dizon, G., & Tang, D. (2017). Comparing the efficacy of digital flashcards versus paper flashcards to improve receptive and productive L2 vocabulary. *The EuroCALL Review*, *25*(1), 3. https://doi.org/10.4995/eurocall.2017.6964

Dóczi, B., & Kormos, J. (2015). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford University Press.

Dörnyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667343

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Lawrence Erlbaum Associates Inc. https://doi.org/10.4324/9781410613349

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.

Dörnyei, Z. (2015). *The psychology of the language learner revisited*. Routledge.

Dörnyei, Z., & Otto, I. (1998). Motivation in action: A process model of L2 motivation. *Working Papers in Applied Linguistics*, *4*, 43–69. http://eprints.nottingham.ac.uk/39/

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Blackwell.

Dörnyei, Z., & Ushioda, E. (2011). *Teaching and researching motivation*. Pearson Education.

Dunn, L., & Dunn, L. (2007). *Peabody picture vocabulary test*. American Guidance Service.

Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*, 1–13. https://doi.org/https://doi.org/10.1016/j.system.2019.102166

Durrant, P. L., Siyanova-Chanturia, A., Kremmel, B., & Sonbul, S. (2022). *Research methods in vocabulary studies* (the). John Benjamins Publishing Company. https://doi.org/10.1075/rmal.2

Dustin, D. S. (1971). Some effects of exam frequency. *The Psychological Record*, *21*(3), 409–414. https://doi.org/10.1007/bf03394033

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research*, *16*(2), 227–252. https://doi.org/10.1177/1362168811431377

Eldridge, J., & Neufield, S. (2009). The graded reader is dead, long live the electronic reader. *Reading*, *9*(2), 224–244. http://www.readingmatrix.com/articles/sept_2009/eldridge_neufeld.pdf

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253–272. https://doi.org/10.1177/0265532212459028

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, *24*(2), 174–187.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. In *Language Learning* (Vol. 43).

Ellis, R. (1999). *Learning a second language through interaction* (Vol. 17). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.17

Ellis, R. (2008). Individual differences in second language learning. *The Handbook of Applied Linguistics*, 525–551. https://doi.org/10.1002/9780470757000.ch21

Ellis, R. (2010). Second language acquisition, teacher education and language pedagogy. *Language Teaching*, *43*(2), 182–201. https://doi.org/10.1017/S0261444809990139

Elyas, T., & Picard, M. (2018). A brief history of English and English teaching in Saudi Arabia. In C. Moskovsky & M. Picard (Eds.), *English as a foreign language in Saudi Arabia: New insights into teaching and learning English* (pp. 70–84). Routledge.

Fan, M. Y. (2003). Frequency of use, perceived usefulness, and actual usefulness of second language vocabulary strategies: A study of Hong Kong learners. *The Modern Language Journal*, *87*(2), 222–241. https://doi.org/10.1111/1540-4781.00187

Fawaaz, N. (2023). Six million students return to school today. *Al Arabiya*. https://www.alarabiya.net/saudi-today/2023/08/20/-6-ملايين-طالب-وطالبة-يعودون-للدراسة-اليوم-

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing. *Studies in Second Language Acquisition*, *42*(3), 499–523. https://doi.org/10.1017/S0272263119000494

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE. https://doi.org/10.5860/CHOICE.50-2114

Fitzpatrick, T. (2012). Tracking the changes: Vocabulary acquisition in the study abroad context. *The Language Learning Journal*, *40*(1), 81–98. https://doi.org/10.1080/09571736.2012.658227

Folse, K. S. (2004). Myths about teaching and learning second language vocabulary: What recent research says. *TESL Reporter*, *37*(2), 1–13.

Fontecha, A. F., & Gallego, M. T. (2012). The role of motivarion and age in vocabulary knowledge. *Vigo International Journal of Applied Linguistics*, *9*(1), 39–62.

Frumuselu, A. D., De Maeyer, S., Donche, V., & Colon Plana, M. del M. G. (2015). Television series inside the EFL classroom: Bridging the gap between teaching and learning informal language through subtitles. *Linguistics and Education*, *32*, 107–117. https://doi.org/10.1016/j.linged.2015.10.001

Gallego, M. T., & Llach, M. D. P. A. (2009). Exploring the increase of receptive vocabulary knowledge in the foreign language: A longitudinal study. *International Journal of English Studies (IJES)*, *9*(1), 113–133. https://doi.org/10.6018/ijes.9.1.90681

García Botero, G., Botero Restrepo, M. A., Zhu, C., & Questier, F. (2021). Complementing in-class language learning with voluntary out-of-class MALL: Does training in self-regulation and scaffolding make a difference? *Computer Assisted Language Learning*, *34*(8), 1013–1039. https://doi.org/10.1080/09588221.2019.1650780

García Botero, G., Questier, F., & Zhu, C. (2019). Self-directed language learning in a mobile-assisted, out-of-class context: Do students walk the talk? *Computer Assisted Language Learning*, *32*(1–2), 71–97. https://doi.org/10.1080/09588221.2018.1485707

Gardner, R. C., Lalonde, R. N., & Moorcroft, R. (1985). Second language learning : Correlational and experimental considerations. *Language Learning*, *35*(2), 207–227.

Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second-language learning*. Newbury House Publishers.

Gass, S. M., & Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 175–199). Erlbaum.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course*. Routledge. https://doi.org/10.2307/416225

GASTAT. (2018). *Bulletin of household culture and entertainment survey*. https://www.stats.gov.sa/sites/default/files/bulletin_of_culture_and_household_entertainment_survey_2018en_0.pdf

Gesa, F., & Miralpeix, I. (2023). Extensive viewing as additional input for foreign language vocabulary learning: A longitudinal study in secondary school. *Language Teaching Research*, 136216882311694. https://doi.org/10.1177/13621688231169451

Gillet, N., Morin, A. J. S., & Reeve, J. (2017). Stability, change, and implications of students' motivation profiles: A latent transition analysis. *Contemporary Educational Psychology*, *51*(August), 222–239. https://doi.org/10.1016/j.cedpsych.2017.08.006

Godwin-Jones, R. (2018). Chasing the butterfly effect: Informal language learning online as a complex system. *Language Learning & Technology*, *22*(2). https://doi.org/10125/44643

Gokcora, D., & DePaulo, D. (2018). Frequent quizzes and student improvement of reading: A pilot study in a community college setting. *SAGE Open*, *8*(2). https://doi.org/10.1177/2158244018782580

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? *ITL - International Journal of Applied Linguistics*, *166*(1), 94–126. https://doi.org/10.1075/itl.166.1.03fer

González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

González-fernández, B., & Schmitt, N. (2020). Word Knowledge: Exploring the Relationships and Order of Acquisition of Vocabulary Knowledge Components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/APPLIN/AMY057

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*(4), 341–363. https://doi.org/10.1093/applin/11.4.341

Green, J. M., & Oxford, R. (1995). A closer look at learning strategies, L2 proficiency, and gender. *TESOL Quarterly*, *29*(2), 261–297.

Griffiths, C. (2020). Language learning strategies: Is the baby still in the bathwater? *Applied Linguistics*, *41*(4), 607–611. https://doi.org/10.1093/applin/amy024

Gu, Y., & Johnson, R. K. (1996). Vocabulary learning strategies and language learning outcomes. *Language Learning*, *46*(4), 643–679. https://doi.org/10.1111/j.1467-1770.1996.tb01355.x

Gurung, R. A. R., & Burns, K. (2019). Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology*, *33*(5), 732–743. https://doi.org/10.1002/acp.3507

Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, *38*(4), 558–579. https://doi.org/10.1177/0265532220979562

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

Hattie, J. A. (1987). Identifying the salient facets of a model of student learning: A synthesis of meta-analyses. *International Journal of Educational Research*, *11*, 187–212.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. https://doi.org/10.1017/S0272263199002089

Hiver, P., Mercer, S., & Al-Hoorie, A. H. (2020). Student engagement in the language classroom. In *Student Engagement in the Language Classroom*. https://doi.org/10.21832/HIVER3606

Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, *21*, 112–126. https://doi.org/10.1017/s0267190501000071

Hu, H. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement load hypothesis versus technique feature analysis. *System*, *56*, 28–39. https://doi.org/10.1016/j.system.2015.11.001

Huang, B., Shawn Chang, Y. H., Niu, L., & Zhi, M. (2018). Examining the effects of socio-economic status and language input on adolescent English learners' speech production outcomes. *System*, *73*, 27–36. https://doi.org/10.1016/j.system.2017.07.004

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, *21*(2), 181–193. https://doi.org/DOI: 10.1017/S0272263199002028

Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 258–286). Cambridge University Press.

Hulstijn, J. H. (2003). Incidental and Intentional Learning. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 349–381). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756492.ch12

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*(3), 327. https://doi.org/10.2307/329439

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Jabbari, N., & Eslami, Z. R. (2023). Negotiations for meaning in the context of a massively multiplayer online role-playing game. *Language Learning & Technology*, *27*(1), 1–28. http://hdl.handle.net/10125/73517

Janebi Enayat, M., & Haghighatpasand, M. (2019). Exploiting adventure video games for second language vocabulary recall: a mixed-methods study. *Innovation in Language Learning and Teaching*, *13*(1), 61–75. https://doi.org/10.1080/17501229.2017.1359276

Jenifer, L.-H. (2016). A guide to doing statistics in second language research using SPSS and R. In *Applied Linguistics*. Routledge.

Jodai, H., Zafarghandi, A. M. V., & Tous, M. D. (2014). Motivation, integrativeness, organizational influence, anxiety, and English achievement. *Glottotheory*, *4*(2), 2–25. https://doi.org/10.1524/glot.2013.0012

Johnson, W. L. (2007). Serious use of a serious game for language learning. *Frontiers in Artificial Intelligence and Applications*, *158*, 67–74.

Johnson, W. L., Vilhjalmsson, H., & Marsella, S. (2005). Serious games for language learning: How much game, how much AI? *Frontiers in Artificial Intelligence and Applications*, *125*, 306–313.

Jordan, G., & Gray, H. (2019). We need to talk about coursebooks. *ELT Journal*, *73*(4), 438–446. https://doi.org/10.1093/elt/ccz038

Jurczyk, V., Fröber, K., & Dreisbach, G. (2019). Increasing reward prospect motivates switching to the more difficult task. *Motivation Science*, *5*(4), 295–313. https://doi.org/10.1037/mot0000119

Kanayama, K., & Kasahara, K. (2018). The indirect effects of testing: Can poor performance in a vocabulary quiz lead to long-term L2 vocabulary retention? *Vocabulary Learning and Instruction*, *7*(1), 1–13. https://doi.org/10.7820/vli.v07.1.kanayama.kasahara

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal*, *43*(1), 53–67. https://doi.org/10.1177/0033688212439359

Karpicke, J., & Roediger, H. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Karpicke, J., & Roediger, H. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research*, *12*(3), 365–386. https://doi.org/10.1177/1362168808089922

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology*, *25*(6), 427–436. https://doi.org/10.1037/h0074468

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education*, *8*(1), 13–24. https://doi.org/10.1016/j.iheduc.2004.12.001

Kika, F. M., McLaughlin, T. F., & Dixon, J. (1992). Effects of frequent testing of secondary algebra students. *Journal of Educational Research*, *85*(3), 159–162. https://doi.org/10.1080/00220671.1992.9944432

Kim, S. K., & Webb, S. (2022). Individual difference factors for second language vocabulary. In S. Li, P. Hiver, & M. Papi (Eds.), *The Routledge handbook of second language acquisition and individual differences* (pp. 282–293). Routledge.

Kim, Y. (2008). The contribution of collaborative and individual tasks to the acquisition of L2 vocabulary. *Modern Language Journal*, *92*(1), 114–130. https://doi.org/10.1111/j.1540-4781.2008.00690.x

Kojic-Sabo, I., & Lightbown, P. M. (1999). Students' approaches to vocabulary learning and their relationship to success. *The Modern Language Journal*, *83*(2), 176–192. https://doi.org/https://doi.org/10.1111/0026-7902.00014

Koolstra, C. M., & Beentjes, J. W. J. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development*, *47*(1), 51–60. https://doi.org/10.1007/BF02299476

Kormos, J., & Csizér, K. (2014). The interaction of motivation, self-regulatory strategies, and autonomous learning behavior in different learner groups. *TESOL Quarterly*, *48*(2), 275–299. https://doi.org/10.1002/tesq.129

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism*, *11*(2), 261–271. https://doi.org/10.1017/S1366728908003416

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297 – 1317. https://doi.org/https://doi.org/10.1002/acp.1537

Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal*, *73*(4), 440. https://doi.org/10.2307/326879

Krashen, S. (2004). *The power of reading*. Libraries Unlimited.

Kremmel, B. (2021). Selling the (word) family silver? A response to Webb's "lemma dilemma." *Studies in Second Language Acquisition*, *43*(5), 962–964. https://doi.org/10.1017/S0272263121000693

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Kruk, M., Pawlak, M., & Zawodniak, J. (2021). Another look at boredom in language instruction: The role of the predictable and the unexpected. *Studies in Second Language Learning and Teaching*, *11*(1), 15–40. https://doi.org/10.14746/ssllt.2021.11.1.2

Kusurkar, R. A., Croiset, G., Galindo-Garré, F., & Ten Cate, O. (2013). Motivational profiles of medical students: Association with study effort, academic performance and exhaustion. *BMC Medical Education*, *13*(1). https://doi.org/10.1186/1472-6920-13-87

Kyndt, E., Dochy, F., Struyven, K., & Cascallar, E. (2011). The direct and indirect effect of motivation for learning on students' approaches to learning through the perceptions of workload and task complexity. *Higher Education Research and Development*, *30*(2), 135–150. https://doi.org/10.1080/07294360.2010.501329

Lachman, S. J., & Forsberg, L. K. (1981). Word recognition and word recall. *Psychological Reports*, *49*(1), 163–170. https://doi.org/10.2466/pr0.1981.49.1.163

LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46–77). Routledge.

Lai, C., Zhu, W., & Gong, G. (2015). Understanding the quality of out-of-class english learning. *TESOL Quarterly*, *49*(2), 278–308. https://doi.org/10.1002/tesq.171

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21–33. https://doi.org/10.1177/003368829402500202

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255–271. https://doi.org/10.1093/applin/19.2.255

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? *Canadian Modern Language Review*, *59*(4), 565–585.

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, *5*, 223–250.

Laufer, B. (2006). Comparing focus on form and focus on forms in second-language vocabulary learning. *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes*, *63*(1), 149–166. https://doi.org/10.1353/cml.2006.0047

Laufer, B. (2009). Second language vocabulary acquisition from language input and from form-focused activities. *Language Teaching*, *42*(2), 341–354. https://doi.org/10.1017/S0261444809005771

Laufer, B. (2010). Form-focused instruction in second language vocabulary learning. In R. Chacón-Beltrán, C. Abello-Contesse, & M. del M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (Issue May, pp. 15–27). Multilingual Matters. https://doi.org/10.21832/9781847692900-003

Laufer, B. (2021). Lemmas, flemmas, word families, and common sense. *Studies in Second Language Acquisition*, *43*(5), 965–968. https://doi.org/10.1017/S0272263121000656

Laufer, B., & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *Modern Language Journal*, *101*(4), 729–741. https://doi.org/10.1111/modl.12431

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

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Laufer, B., & Levitzky-Aviad, T. (2018). Loanword proportion in vocabulary size tests. *ITL - International Journal of Applied Linguistics*, *169*(1), 95–114. https://doi.org/10.1075/itl.00008.lau

Laufer, B., & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly*, *13*(3), 202–217. https://doi.org/10.1080/15434303.2016.1210611

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 36–55. https://doi.org/10.1191/026553299672614616

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365–391. https://doi.org/10.1111/0023-8333.00046

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30. http://nflrc.hawaii.edu/rfl

Laufer, B., & Rozovski-Roitblat, B. (2015). Retention of new words: Quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research*, *19*(6), 687–711. https://doi.org/10.1177/1362168814559797

Lee, J. H., Ahn, J. J., & Lee, H. (2022). The role of motivation and vocabulary learning strategies in L2 vocabulary knowledge: A structural equation modeling analysis. *Studies in Second Language Learning and Teaching*, *12*(3), 435–458. https://doi.org/10.14746/ssllt.2022.12.3.5

Lee, J. H., & Lee, H. (2022). Teachers' verbal lexical explanation for second language vocabulary learning: A meta-analysis. *Language Learning*. https://doi.org/10.1111/lang.12493

Lee, S. (2020). Examining the roles of aptitude, motivation, strategy use, language processing experience, and gender in the development of the breadth and depth of EFL learners' vocabulary knowledge. *SAGE Open*, *10*(4). https://doi.org/10.1177/2158244020977883

Lefever, S. (2010). English skills of young learners in Iceland "I started talking English when I was 4 years. It just bang...just fall into me.". *Ráðstefnurit Netlu – Menntakvika 2010*, *December*, 1–17.

Li, M., & Hennebry-Leung, M. (2022). Effects of monolingual and bilingual subtitles on L2 vocabulary acquisition. *IRAL - International Review of Applied Linguistics in Language Teaching*, 1–28. https://doi.org/10.1515/iral-2022-0034

Li, M., & Kirby, J. R. (2015). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, *36*(5), 611–634. https://doi.org/10.1093/applin/amu007

Li, S., Hiver, P., & Papi, M. (2022). *The Routledge handbook of second language acquisition and individual differences* (S. Li, P. Hiver, & M. Papi, Eds.). Routledge. https://doi.org/10.4324/9781003270546

Li, Y., & Zhang, X. (2019). L2 Vocabulary knowledge and L2 listening comprehension: A structural equation model. *Canadian Journal of Applied Linguistics*, *22*(1), 85–102. https://doi.org/10.7202/1060907ar

Li, Z., & Bonk, C. J. (2023). Self-directed language learning with Duolingo in an out-of-class context. *Computer Assisted Language Learning*, 1–23. https://doi.org/10.1080/09588221.2023.2206874

Lightbown, P. M., & Spada, N. (2020). Teaching and learning L2 in the classroom: It's about time. *Language Teaching*, *53*(4), 422–432. https://doi.org/10.1017/S0261444819000454

Linck, J. A. (2016). Analyzing individual differences in second language research: The benefits of mixed effects models. In G. Granena, D. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 105–128). https://doi.org/10.1075/bpa.3.06lin

Liu, M. (2007). Chinese students' motivation to learn English at the tertiary level. *Asian EFL Journal*, *9*(1), 86–96. http://www.asian-efl-journal.com/March_2007_EBook.pdf

Liu, M., & Oga-Baldwin, W. L. Q. (2022). Motivational profiles of learners of multiple foreign languages: A self-determination theory perspective. *System*, *106*, 102762. https://doi.org/10.1016/j.system.2022.102762

Long, M. H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input1. *Applied Linguistics*, *4*(2), 126–141. https://doi.org/10.1093/applin/4.2.126

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsh (Eds.), *Foreign Language Research in Cross-cultural Perspective* (pp. 39–52). John Benjamins.

Lwo, L., & Chia-Tzu Lin, M. (2012). The effects of captions in teenagers' multimedia L2 learning. *ReCALL*, *24*(2), 188–208. https://doi.org/10.1017/S0958344012000067

Macaro, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *Modern Language Journal*, *90*(3), 320–337. https://doi.org/10.1111/j.1540-4781.2006.00425.x

Mahboob, A., & Elyas, T. (2014). English in the Kingdom of Saudi Arabia. *World Englishes*, *33*(1), 128–142. https://doi.org/10.1111/weng.12073

Mandler, G. (2008). Familiarity breeds attempts: A critical review of dual-process theories of recognition. *Perspectives on Psychological Science*, *3*(5), 390–399. https://doi.org/10.1111/j.1745-6924.2008.00087.x

Mantle-Bromley, C. (1995). Positive attitudes and realistic beliefs: Links to proficiency. *The Modern Language Journal*, *79*(3), 372–386. https://doi.org/10.1111/j.1540-4781.1995.tb01114.x

Markham, P. L., Peter, L. A., & McCarthy, T. J. (2001). The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign Language Annals*, *34*(5), 439–445. https://doi.org/10.1111/j.1944-9720.2001.tb02083.x

Masrai, A., & Milton, J. (2015). An investigation of the relationship between L1 lexical translation equivalence and L2 vocabulary acquisition. *International Journal of English Linguistics*, *5*(2), 1–7. https://doi.org/10.5539/ijel.v5n2p1

Masrai, A., & Milton, J. (2018a). The role of informal learning activities in improving L2 lexical access and acquisition in L1 Arabic speakers learning EFL. *The Language Learning Journal*, *46*(5), 594–604. https://doi.org/10.1080/09571736.2018.1520655

Masrai, A., & Milton, J. (2018b). The role of informal learning activities in improving L2 lexical access and acquisition in L1 Arabic speakers learning EFL. *The Language Learning Journal*, *46*(5), 594–604. https://doi.org/10.1080/09571736.2018.1520655

Mayer, R. E. (2001). *Multimedia Learning*. Cambridge University Press. https://doi.org/10.1007/978-1-4419-1428-6_285

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amw050

Mclean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction*, *3*(2), 47–55. https://doi.org/10.7820/vli.v03.2.mclean.et.al

McLean, S., Hogg, N., & Rush, T. W. (2013). Vocabulary learning through an online computerized flashcard site. *Jaltcalljournal*, *9*(1), 79–98.

McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. *Shiken*, *19*(2), 1–11. http://teval.jalt.org/sites/teval.jalt.org/files/Shiken_19-02_Complete-1.pdf#page=26

Mclean, S., Kramer, B., & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4*(1), 1–10. https://doi.org/10.7820/vli.v04.1.mclean.et.al

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389–411. https://doi.org/10.1177/0265532219898380

Meara, P. (1980). *Meara 1980 Vocabulary acquisition: a neglected aspect of language learning. 1*, 1–29.

Meara, P. (1992). *EFL vocabulary tests*. University College, Swansea: Centre for Applied Language Studies.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, *18*(4), 393–407. https://doi.org/10.1191/0267658302sr211xx

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19–30. https://doi.org/10.1016/S0346-251X(99)00058-5

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied Linguistics in Society.* (pp. 80–87). CIL.

Meara, P., & Jones, G. (1990). *The eurocentres vocabulary size test 10k*. Eurocentres.

Medina, S. L. (1993). The effects of music upon second language vocabulary acquisition. *National Network for Early Language Learning*, *6*, 1–8. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title#0%5Cnhttp://eric.ed.gov/?id=ED352834

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary description, acquisition and pedagogy* (pp. 84–102). Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Michael, J. (1991). A behavioral perspective on college teaching. *The Behavior Analyst*, *14*(2), 229–239. https://doi.org/10.1007/bf03392578

Milton, J. (2006). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, *16*(2), 187–205. https://doi.org/10.1017/S0959269506002420

Milton, J. (2008). Vocabulary uptake from informal learning tasks. *Language Learning Journal*, *36*(2), 227–237. https://doi.org/10.1080/09571730802390742

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters. https://doi.org/10.21832/9781847692092

Milton, J., & Fitzpatrick, T. (2014). *Dimensions of vocabulary knowledge* (J. Milton & T. Fitzpatrick, Eds.). Palgrave Macmillan.

Milton, J., & Hopwood, O. (2022). *Vocabulary in the foreign language curriculum*. Routledge. https://doi.org/10.4324/9781003278771

Milton, J., & Meara, P. (1998). Are the British really bad at learning foreign languages? *Language Learning Journal*, *18*(1), 68–76. https://doi.org/10.1080/09571739885200291

Milton, J., & Vassiliu, P. (2000). Frequency and the lexis of low level EFL texts. In K. Nicolaidis & M. Mattheoudakis (Eds.), *13th Symposium on Theoretical and Applied Linguistics* (Vol. 13, pp. 444–455). e-issn: 2529-1114

Ministry of Education. (2023). *Guide to school curriculum*.

Miralpeix, I., & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *IRAL - International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. https://doi.org/10.1515/iral-2017-0016

Mizumoto, A., & Takeuchi, O. (2011). Adaptation and validation of self-regulating capacity in vocabulary learning scale. *Applied Linguistics*, *33*(1), 83–91. https://doi.org/10.1093/applin/amr044

Mohsen, M. A. (2016). The use of computer-based simulation to aid comprehension and incidental vocabulary learning. *Journal of Educational Computing Research*, *54*(6), 863–884. https://doi.org/10.1177/0735633116639954

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: The effect of two enhancement techniques. *Computer Assisted Language Learning*, *31*(1–2), 1–26. https://doi.org/10.1080/09588221.2017.1375960

Morgan, J., & Rinvolucri, M. (2004). *Vocabulary*. Open University Press.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Moskovsky, C., Alrabai, F., Paolini, S., & Ratcheva, S. (2013). The effects of teachers' motivational strategies on learners' motivation: A controlled investigation of second language acquisition. *Language Learning*, *63*(1), 34–62. https://doi.org/10.1111/j.1467-9922.2012.00717.x

Moskovsky, C., & Picard, M. (2018). *English as a foreign language in Saudi Arabia: New insights into teaching and learning English*. Routledge. www.routledge.com/Routledge-

Muñoz, C. (2017). The role of age and proficiency in subtitle reading. An eye-tracking study. *System*, *67*, 77–86. https://doi.org/10.1016/j.system.2017.04.015

Muñoz, C., Pujadas, G., & Pattemore, A. (2021). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research*. https://doi.org/10.1177/02676583211015797

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. In *American Educational Research Journal* (Vol. 24, Issue 2). https://doi.org/10.3102/00028312024002237

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Quarterly*, *49*(1), 6–37. https://doi.org/10.1002/tesq.157

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, *20*(1), 3–20. https://doi.org/10.1017/S0958344008000219

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: a critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*(1), 17–38. https://doi.org/10.1080/09588221.2010.520675

Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning. *IRAL - International Review of Applied Linguistics in Language Teaching*, *54*(3), 257–289. https://doi.org/10.1515/iral-2015-0022

Nakata, T. (2020). Learning words with flash cards and word cards. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 304–319). Routledge.

Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*(2), 233–260. https://doi.org/10.1177/0267658320927764

Nakata, T., Tada, S., Mclean, S., & Kim, Y. A. (2021). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly*, *55*(1), 248–270. https://doi.org/10.1002/tesq.596

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning? *Studies in Second Language Acquisition*, *38*(3), 523–552. https://doi.org/10.1017/S0272263115000236

Nation, P. (1983). Testing and teaching vocabulary 5(1), 12–25. *Guidelines*, *5*(1), 12–25.

Nation, P. (1990). *Teaching and Learning Vocabulary*. Heinle and Heinle.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, P. (2007). The four strands. *Innovation in Language Learning and Teaching*, *1*(1), 2–13. https://doi.org/10.2167/illt039.0

Nation, P. (2011). Research into practice: Vocabulary. *Language Teaching*, *44*(4), 529–539. https://doi.org/10.1017/s0261444811000267

Nation, P. (2012). The vocabulary size test. *The Language Teacher*, *31*(7), 9–13. https://doi.org/10.1002/9781405198431.wbeal1270

Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.208

Nation, P. (2020). Is it worth teaching vocabulary? *TESOL Journal*, *12*(4). https://doi.org/10.1002/tesj.564

Nation, P. (2021). Thoughts on word families. *Studies in Second Language Acquisition*, *43*(5), 969–972. https://doi.org/10.1017/S027226312100067X

Nation, P. (2022). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*. https://doi.org/10.1177/0265532209340194

Nation, P., & Macalister, J. (2020). *Teaching ESL/EFL reading and writing*. Routledge.

Nation, P., & Waring, R. (2020). *Teaching extensive reading in another language*. Routledge. https://doi.org/10.1002/9781118784235.eelt0564

Nation, P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Cengage Learning.

Nation, P., & Yamamoto, A. (2012). Applying the four strands to language learning. *International Journal of Innovation in English Language Teaching and Research*, *1*(2), 167–181.

National Center for Education Statistics. (2020). *National assessment of educational progress*.

National Literacy Trust. (2023). *Children and young people's reading in 2023*.

Ngoc Yen, D. T. (2020). Vietnamese non-English Major EFL university students' receptive knowledge of the most frequent English words. *VNU Journal of Foreign Studies*, *36*(3), 1–11. https://doi.org/10.25073/2525-2445/vnufs.4553

Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. https://doi.org/10.1177/0033688210390264

Nikoopour, J., & Kazemi, A. (2014). Vocabulary Learning through digitized & non-digitized flashcards delivery. *Procedia - Social and Behavioral Sciences*, *98*, 1366–1373. https://doi.org/10.1016/j.sbspro.2014.03.554

Noels, K. A., Clément, R., & Pelletier, L. G. (1999). Perceptions of teachers' communicative style and students' intrinsic and extrinsic motivation. *Modern Language Journal*, *83*(1), 23–34. https://doi.org/10.1111/0026-7902.00003

Noels, K. A., Vargas Lascano, D. I., & Saumure, K. (2019). The development of self-determination across the language course. *Studies in Second Language Acquisition*, *41*(4), 821–851. https://doi.org/10.1017/S0272263118000189

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, *18*(2), 161–175. https://doi.org/10.1016/S0889-4906(98)00005-2

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge University Press.

O'Malley, J. M., Chamot, A. U., Stewner-Manzanares, G., Kupper, L., & Russo, R. P. (1985). Learning strategies used by beginning and intermediate ESL students. *Language Learning*, *35*(1), 21–46. https://doi.org/10.1111/j.1467-1770.1985.tb01013.x

Ortega, L. (2019). SLA and the study of equitable multilingualism. *Modern Language Journal*, *103*, 23–38. https://doi.org/10.1111/modl.12525

Oxford, R. L. (1990). *Learning strategies: What every teacher should know* (p. 136).

Oxford, R. L. (2011). *Teaching and researching language learning strategies*. Pearson Education.

Oxford, R. L. (2017). *Teaching and researching language learning strategies*. Routledge. https://doi.org/10.4324/9781315838816

Ozturk, M. (2012). Vocabulary growth of the advanced EFL learner. *The Language Learning Journal*, *43*(1), 94–109. https://doi.org/10.1080/09571736.2012.708053

Palmer, E. L. (1974). Frequency of tests and general subject-area mastery. *Psychological Reports*, *35*, 422.

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition*, *41*(4), 745–768. https://doi.org/10.1017/S0272263119000020

Pawlak, M. (2021). Investigating language learning strategies: Prospects, pitfalls and challenges. *Language Teaching Research*, *25*(5), 817–835. https://doi.org/10.1177/1362168819876156

Pawlak, M., Zawodniak, J., & Kruk, M. (2020). *Boredom in the foreign language classroom*. Springer International Publishing. https://doi.org/10.1007/978-3-030-50769-5

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*, *38*(1), 97–130. https://doi.org/10.1017/S0272263115000224

Pellicer-Sánchez, A. (2019). Examining second language vocabulary growth: Replications of Schmitt (1998) and Webb & Chang (2012). *Language Teaching*, *52*(4), 512–523. https://doi.org/10.1017/S026144481800037X

Peters, E. (2018). The effect of out-of-class exposure to English language media on learners' vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, *169*(1), 142–168. https://doi.org/10.1075/itl.00010.pet

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, *53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System*, *63*, 134–148. https://doi.org/10.1016/j.system.2016.10.002

Peters, E., Velghe, T., & Van Rompaey, T. (2019). The VocabLab tests. *ITL - International Journal of Applied Linguistics*, *170*(1), 53–78. https://doi.org/10.1075/itl.17029.pet

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing l2 television and factors that affect learning. *Studies in Second Language Acquisition*, *40*(3), 551–577. https://doi.org/10.1017/S0272263117000407

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28.

Platzer, H. (2020). The role of Quizlet in vocabulary acquisition. *Electronic Journal of Foreign Language Teaching*, *17*(2), 421–438.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*(4), 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2016). The n crowd: Sampling practices, internal validity, and generalizability in L2 research. In *Presentation at University College London*.

Politzer, R. L. (1978). Errors of English speakers of German as perceived and evaluated by German natives. *The Modern Language Journal*, *62*(5/6), 253. https://doi.org/10.2307/324907

Puimège, E., & Peters, E. (2019). Learning formulaic sequences through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, 1–25. https://doi.org/10.1017/s027226311900055x

Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: A study of L2 vocabulary learning by adolescents. *Language Learning Journal*, *47*(4), 479–496. https://doi.org/10.1080/09571736.2019.1616806

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge University Press.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282–308. https://doi.org/10.3138/cmlr.56.2.282

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52. https://doi.org/10.1191/0265532204lt273oa

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Read, J. (2019). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 545–560). Routledge.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1–32. https://doi.org/10.1177/026553220101800101

Rees-Miller, J. (1993). A critical appraisal of learner training: Theoretical bases and teaching implications. *TESOL Quarterly*, *27*(4), 679. https://doi.org/10.2307/3587401

Reinhardt, J. (2019). Social media in second and foreign language teaching and learning: Blogs, wikis, and social networking. *Language Teaching*, *52*(1), 1–39. https://doi.org/10.1017/S0261444818000356

Reinhardt, J., & Thorne, S. (2016). Metaphors for digital games and language learning. In F. Farr & L. Murray (Eds.), *The Routledge handbook of language learning and technology* (pp. 415–430). https://doi.org/10.4324/9781315657899-45

Reynolds, B. L. (2023). *Vocabulary learning in the wild*. Springer. https://doi.org/10.1007/978-981-99-1490-6

Reynolds, B. L., Cui, Y., Kao, C. W., & Thomas, N. (2022). Vocabulary acquisition through viewing captioned and subtitled video: A scoping review and meta-analysis. *Systems*, *10*(5), 1–20. https://doi.org/10.3390/systems10050133

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77. https://doi.org/10.2307/3585941

Richards, J. C. (2006). *Communicative language teaching today*. Cambridge University Press.

Richards, J. C. (2015). The changing face of language learning: Learning beyond the classroom. *RELC Journal*, *46*(1), 5–22. https://doi.org/10.1177/0033688214561621

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge University Press. https://doi.org/10.2307/3588247

Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language*, *108*(2019), 1–17. https://doi.org/10.1016/j.jml.2019.104029

Robles-García, P. (2022). Receptive vocabulary knowledge in L2 learners of Spanish: The role of high-frequency words. *Foreign Language Annals*. https://doi.org/10.1111/flan.12630

Rodrigo, V., Krashen, S., & Gribbons, B. (2004). The effectiveness of two comprehensible-input approaches to foreign language instruction at the intermediate level. *System*, *32*(1), 53–60. https://doi.org/10.1016/j.system.2003.08.003

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Romanko, R. (2017). Measuring the vocabulary burden of popular English songs. *Vocabulary Learning and Instruction*, *6*(2), 71–78. https://doi.org/10.7820/vli.v06.2.romanko

Rose, H., Briggs, J. G., Boggs, J. A., Sergio, L., & Ivanova-Slavianskaia, N. (2018). A systematic review of language learner strategy research in the face of self-regulation. *System*, *72*, 151–163. https://doi.org/10.1016/j.system.2017.12.002

Ross, C. C., & Henry, L. K. (1939). The relation between frequency of testing and progress in learning psychology. *Journal of Educational Psychology*, *30*(8), 604–611. https://doi.org/10.1037/h0055717

Rubin, J. (1975). What the "good language learner" can teach us. *TESOL Quarterly*, *9*(1), 41–51. https://doi.org/10.2307/3586011

Rubin, J. (1981). Study of cognitive processes in second language learning. *Applied Linguistics*, *II*(2), 117–131. https://doi.org/10.1093/applin/II.2.117

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Publications.

Sage, K., Krebs, B., & Grove, R. (2019). Flip, slide, or swipe? Learning outcomes from paper, computer, and tablet flashcards. *Technology, Knowledge and Learning*, *24*(3), 461–482. https://doi.org/10.1007/s10758-017-9345-9

Sage, K., Piazzini, M., Charles, J., Iv, D., & Ewing, S. (2020). Flip it or click it : Equivalent learning of vocabulary from paper , laptop , and smartphone flashcards. *Journal of Educational Technology Systems*. https://doi.org/10.1177/0047239520943647

Sanaoui, R. (1995). Adult learners' approaches to learning vocabulary in second languages. *The Modern Language Journal*, *79*(1), 15–28.

Sanosi, A. B. (2018). The effect of Quizlet on vocabulary acquisition. *Asian Journal of Education and E-Learning*, *06*(04), 71–77.

Sato, M., & Loewen, S. (2019a). Do teachers care about research? The research-pedagogy dialogue. *ELT Journal*, *73*(1), 1–10. https://doi.org/10.1093/elt/ccy048

Sato, M., & Loewen, S. (2019b). Methodological strengths, challenges, and joys of classroom-based quasi-experimental research: Metacognitive instruction and corrective feedback. In R. DeKeyser & G. P. Botana (Eds.), *Doing SLA research with implications for the classroom: Reconciling methodological demands and pedagogical applicability* (pp. 31–54). John Benjamins.

Sato, M., & Loewen, S. (2022). The research–practice dialogue in second language learning and teaching: Past, present, and future. *Modern Language Journal*, *106*(3), 509–527. https://doi.org/10.1111/modl.12791

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy.* (pp. 199–227). Cambridge University Press.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*(2), 281–317. https://doi.org/10.1111/1467-9922.00042

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261–274. https://doi.org/10.1017/S0261444819000053

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, *50*(2), 212–226. https://doi.org/10.1017/S0261444815000075

Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing knowledge-based vocabulary lists (KVL). *TESOL Journal*, *12*(e622). https://doi.org/10.1002/tesj.622

Schmitt, N., Nation, P., & Kremmel, B. (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*, 109–120. https://doi.org/10.1017/S0261444819000326

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/S0261444812000018

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge University Press.

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

Schrank, Z. (2016). An assessment of student perceptions and responses to frequent low-stakes testing in introductory sociology classes. *Teaching Sociology*, *44*(2), 118–127. https://doi.org/10.1177/0092055X15624745

Schunk, D. H., & Green, J. A. (2018). *Handbook of self-regulation of learning and performance*. https://doi.org/10.3102/0034654316628645

Seibert Hanson, A. E., & Brown, C. M. (2019). Enhancing L2 learning through a mobile assisted spaced-repetition tool: an effective but bitter pill? *Computer Assisted Language Learning*, *33*(1–2), 133–155. https://doi.org/10.1080/09588221.2018.1552975

Seker, M. (2016). The use of self-regulation strategies by foreign language learners and its role in language achievement. *Language Teaching Research*, *20*(5), 600–618. https://doi.org/10.1177/1362168815578550

Shadiev, R., Hwang, W. Y., & Huang, Y. M. (2017). Review of research on mobile language learning in authentic environments. *Computer Assisted Language Learning*, *30*(3–4), 284–303. https://doi.org/10.1080/09588221.2017.1308383

Simos, P. G., Sideridis, G. D., Mouzaki, A., Chatzidaki, A., & Tzevelekou, M. (2012). Vocabulary growth in second language among immigrant school-aged children in Greece. *Applied Psycholinguistics*, *35*(3), 621–647. https://doi.org/10.1017/S0142716412000525

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (2018). Understanding formulaic language: A second language acquisition perspective. In *Understanding Formulaic Language: A Second Language Acquisition Perspective*. https://doi.org/10.4324/9781315206615

Siyanova-Chanturia, A., & Webb, S. (2016). Teaching vocabulary in the EFL context. *English Language Teaching Today*, *5*, 227–239. https://doi.org/10.1007/978-3-319-38834-2_16

Skehan, P. (1989). *Individual differences in second language acquisition*. Hodder Education.

Smidt, E., & Hegelheimer, V. (2004). Effects of online academic lectures on ESL listening comprehension, incidental vocabulary acquisition, and strategy use. *Computer Assisted Language Learning*, *17*(5), 517–556. https://doi.org/10.1080/0958822042000319692

Smit, U. (2023). English-medium instruction (EMI). *ELT Journal*, *77*(4), 499–503. https://doi.org/10.1093/elt/ccad018

Sonbul, S., Almusharraf, N., & Abdel Salam El-Dakhs, D. (2022). Examining teachers' awareness of effective vocabulary instructional practices: Voices from Saudi Arabia. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.12441

Sonbul, S., & Schmitt, N. (2009). Direct teaching of vocabulary after reading: Is it worth the effort? *ELT Journal*, *64*(3), 253–260. https://doi.org/10.1093/elt/ccp059

Spada, N. (2015). SLA research and L2 pedagogy: Misapplications and questions of relevance. *Language Teaching*, *48*(1), 69–81. https://doi.org/10.1017/S026144481200050X

Spada, N., & Lightbown, P. M. (2022). In it together: Teachers, researchers, and classroom SLA. *The Modern Language Journal*, *106*(3), 635–650. https://doi.org/10.1111/modl.12792

Spolsky, B. (2000). Anniversary article language motivation revisited. *Applied Linguistics*, *21*(2), 157–169. https://doi.org/10.1093/applin/21.2.157

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Statista. (2022). *Hours of video uploaded to YouTube every minute 2007-2022*. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

Stern, H. H. (1983). *Fundamental concepts in language learning*. Oxford University Press.

Sternberg, RJ. (1987). Most vocabulary is learned from context. In M. MG & C. ME (Eds.), *The nature of vocabulary acquisition* (pp. 89–105). Lawrence Erlbaum Associates Inc.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2024). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*, *41*(1), 89–108. https://doi.org/10.1177/02655322231162853

Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. G. (2021). What the research shows about written receptive vocabulary testing: A reply to Webb. *Studies in Second Language Acquisition*, *43*(2), 462–471. https://doi.org/10.1017/S0272263121000437

Stoeckel, T. (2018). High-Frequency and academic English vocabulary growth among first-year students at UNP. *JISRD*, *9*(9), 15–29.

Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics*, *41*(4), 601–606. https://doi.org/10.1093/applin/amy059

Stoeckel, T., McLean, S., & Nation, P. (2020). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 1–23. https://doi.org/10.1017/S027226312000025X

Stoeckel, T., Reagan, N., & Hann, F. (2012). Extensive reading quizzes and reading attitudes. *TESOL Quarterly*, *46*(1), 187–198.

Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the vocabulary size test to a criterion measure of meaning recall vocabulary knowledge. *System*, *87*, 1–14. https://doi.org/10.1016/j.system.2019.102161

Suk, N. (2017). The effects of extensive reading on reading comprehension, reading rate, and vocabulary acquisition. *Reading Research Quarterly*, *52*(1), 73–89. https://doi.org/10.1002/rrq.152

Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System*, *93*, 102279. https://doi.org/10.1016/j.system.2020.102279

Sundqvist, P. (2009). Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary. In *Faculty of Arts and Education - English: Vol. PhD*. Karlstad University Studies.

Sundqvist, P. (2013). The SSI model: Categorization of digital games in EFL studies. *The European Journal of Applied Linguistics and TEFL*, *2*(1), 89–104. https://link.gale.com/apps/doc/A528960769/AONE?u=anon~cad2c0e1&sid=googleScholar&xid=51462ce7

Sundqvist, P. (2019). Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning and Technology*, *23*(1), 87–113. https://doi.org/10.125/44674

Sundqvist, P. (2024). Extramural English as an individual difference variable in L2 research: Methodology matters. *Annual Review of Applied Linguistics*, 1–13. https://doi.org/10.1017/S0267190524000072

Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in teaching and learning: From theory and research to practice*. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-46048-6

Sundqvist, P., & Wikström, P. (2015). Out-of-school digital gameplay and in-school L2 English vocabulary outcomes. *System*, *51*, 65–76. https://doi.org/10.1016/j.system.2015.04.001

Swanborn, M. S. L., & De Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261–285. https://doi.org/10.3102/00346543069003261

Sylvén, L. K., & Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. *ReCALL*, *24*(3), 302–321. https://doi.org/10.1017/S095834401200016X

Taguchi, E., Takayasu-Maass, M., & Gorsuch, G. J. (2004). Developing reading fluency in EFL : How assisted repeated reading and extensive reading affect fluency development. *Reading in a Foreign Language*, *16*(2), 70–96.

Takač, V. P. (2008). *Vocabulary learning strategies and foreign language acquisition*. Multilingual Matters.

Tanaka, M. (2017). Examining EFL vocabulary learning motivation in a demotivating learning environment. *System*, *65*, 130–138. https://doi.org/10.1016/j.system.2017.01.010

Taris, T. (2000). *A primer in longitudinal data analysis*. SAGE Publications Ltd. https://doi.org/10.4135/9781849208512

Team, R. C. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Tegge, F. (2017). The lexical coverage of popular songs in English language teaching. *System*, *67*, 87–98. https://doi.org/10.1016/j.system.2017.04.016

Teng, M. F. (2022). Incidental L2 vocabulary learning from viewing captioned videos: Effects of learner-related factors. *System*, *105*, 102736. https://doi.org/10.1016/j.system.2022.102736

Teravainen-Goff, A. (2022). Why motivated learners might not engage in language learning: An exploratory interview study of language learners and teachers. *Language Teaching Research*. https://doi.org/10.1177/13621688221135399

Thomas, N. (2020). Incidental L2 vocabulary learning: Recent developments and implications for future research. *Reading in a Foreign Language*, *32*(1), 49–60.

Tremblay, P. F., Goldberg, M. P., & Gardner, R. C. (1995). Trait and state motivation and the acquisition of Hebrew vocabulary. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *27*(3), 356.

Tseng, W. T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics*, *27*(1), 78–102. https://doi.org/10.1093/applin/ami046

Tseng, W. T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*(2), 357–400. https://doi.org/10.1111/j.1467-9922.2008.00444.x

Tuckman, B. W. (1998). Using tests as an incentive to motivate procrastinators to study. *Journal of Experimental Education*, *66*(2), 141–147. https://doi.org/10.1080/00220979809601400

Uchihara, T. (2023). How does the test modality of weekly quizzes influence learning the spoken forms of second language vocabulary? *TESOL Quarterly*, *57*(2), 595–617. https://doi.org/10.1002/tesq.3176

Uchihara, T., & Clenton, J. (2023). The role of spoken vocabulary knowledge in second language speaking proficiency. *Language Learning Journal*, *51*(3), 376–393. https://doi.org/10.1080/09571736.2022.2080856

Uchihara, T., & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *Language Learning Journal*, *47*(1), 64–75. https://doi.org/10.1080/09571736.2016.1191527

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599. https://doi.org/10.1111/lang.12343

Ushioda, E. (2020). *Language learning motivation*. Oxford University Press.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. F. (1992). The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. educational and psychological neasurement. In *Educational and Psychological Measurement* (Vol. 52, Issue 4).

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. https://doi.org/10.1093/applin/ams074

Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., & Lens, W. (2009). Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of Educational Psychology*, *101*(3), 671–688. https://doi.org/10.1037/a0015083

Vu, D. Van, & Peters, E. (2021). Vocabulary in English Language Learning, Teaching, and Testing in Vietnam: A Review. *Education Sciences*, *11*(9), 563. https://doi.org/10.3390/educsci11090563

Walker, E. A., Redfern, A., & Oleson, J. J. (2019). Linear mixed-model analysis to examine longitudinal trajectories in vocabulary depth and breadth in children who are hard of hearing. *Journal of Speech, Language, and Hearing Research*, *62*(3), 525–542. https://doi.org/10.1044/2018_JSLHR-L-ASTM-18-0250

Wang, A., & Pellicer-Sánchez, A. (2022). Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning*. https://doi.org/10.1111/lang.12495

Wang, C., & Sun, T. (2020). Relationship between self-efficacy and language proficiency: A meta-analysis. *System*, *95*. https://doi.org/10.1016/j.system.2020.102366

Wang, F. X. (2008). Motivation and English achievement: An exploratory and confirmatory factor analysis of a new measure for Chinese students of English learning. *North American Journal of Psychology*, *10*(3), 633–646.

Wang, H., & Chen, C. W. (2020). Learning English from youtubers: English L2 learners' self-regulated language learning on YouTube. *Innovation in Language Learning and Teaching*, *14*(4), 333–346. https://doi.org/10.1080/17501229.2019.1607356

Waring, R. (1997). A study of receptive and productive vocabulary learning from word cards. *Studies in Foreign Languages and Literature*, *21*(1), 94–114.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 79–95. https://doi.org/10.1017/S0272263105050023

Webb, S. (2007a). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*(1), 63–81. https://doi.org/10.1177/1362168806072463

Webb, S. (2007b). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, *65*(3), 441–470. https://doi.org/10.3138/cmlr.65.3.441

Webb, S. (2019). The Routledge handbook of vocabulary studies. In *The Routledge Handbook of Vocabulary Studies*. Routledge. https://doi.org/10.4324/9780429291586

Webb, S. (2021a). A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(2), 454–461. https://doi.org/10.1017/S0272263121000449

Webb, S. (2021b). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, *33*(2), 278–293. https://nflrc.hawaii.edu/rfl/

Webb, S. (2021c). The lemma dilemma: How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition*, *43*(5), 941–949. https://doi.org/10.1017/S0272263121000784

Webb, S. (2021d). Word families and lemmas, not a real dilemma. *Studies in Second Language Acquisition*, *43*(5), 973–984. https://doi.org/10.1017/S0272263121000760

Webb, S., & Chang, A. C. S. (2012). Second language vocabulary growth. *RELC Journal*, *43*(1), 113–126. https://doi.org/10.1177/0033688212439367

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Webb, S., & Rodgers, M. P. H. (2009a). The Lexical Coverage of Movies. *Applied Linguistics*, *30*(3), 407–427. https://doi.org/10.1093/applin/amp010

Webb, S., & Rodgers, M. P. H. (2009b). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Webb, S., Uchihara, T., & Yanagisawa, A. (2023). How effective is second language incidental vocabulary learning? A meta-analysis. *Language Teaching*, 1–20. https://doi.org/10.1017/S0261444822000507

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*. https://doi.org/10.1111/modl.12671

Wei, R., & Fan, L. (2022). On-screen texts in audiovisual input for L2 vocabulary learning: A review. *Frontiers in Psychology*, *13*(May), 1–11. https://doi.org/10.3389/fpsyg.2022.904523

Weiner, B. (1992). *Human motivation: Metaphors, theories and research*. SAGE.

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(4), 1039–1048. https://doi.org/10.1037/a0036164

Wen, Z., Skehan, P., Biedroń, A., Li, S., & Sparks, R. L. (2019). Language aptitude: Advancing theory, testing, research and practice. In *Language Aptitude: Advancing Theory, Testing, Research and Practice*. https://doi.org/10.4324/9781315122021

Wenden, A. (1991). *Learner strategies for learner autonomy*. Prentice Hall.

West, Michael. (1953). *A general service list of English words*. Longman, Green & Co.

Wharton, G. (2000). Language learning strategy use of bilingual foreign language learners in Singapore. *Language Learning*, *50*(2), 203–243. https://doi.org/10.1111/0023-8333.00117

Wilkins, D. A. (1972). *Linguistics in language teaching*. Arnold.

Wilkinson, D. (2020). *Effects of word card methodology and testing on vocabulary knowledge and motivation* (Issue August) [Temple University]. https://www.proquest.com/docview/2445495447?pq-origsite=gscholar&fromopenview=true

Winke, P., Gass, S. M., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*(1), 65–86. http://llt.msu.edu/vol14num1/winkegasssydorenko.pdf

Wood, D. (2015). *Fundamentals of formulaic language*. Bloomsbury Academic.

Wu, S., Quentin Dixon, L., Sun, H., & Zhang, P. (2021). Breadth or depth: the role of vocabulary in Chinese English-Language beginning writers' development. *International Journal of Bilingual Education and Bilingualism*, *24*(9), 1356–1372. https://doi.org/10.1080/13670050.2019.1572066

Xodabande, I., Asadi, V., & Valizadeh, M. (2022). Teaching vocabulary items in corpus-based wordlists to university students: Comparing the effectiveness of digital and paper-based flashcards. *Journal of China Computer-Assisted Language Learning*, *0*(0). https://doi.org/10.1515/jccall-2022-0016

Xodabande, I., Iravi, Y., Mansouri, B., & Matinparsa, H. (2022). Teaching academic words with digital flashcards: Investigating the effectiveness of mobile-assisted vocabulary learning for university students. *Frontiers in Psychology*, *13*(June). https://doi.org/10.3389/fpsyg.2022.893821

Xodabande, I., Pourhassan, A., & Valizadeh, M. (2022). Self-directed learning of core vocabulary in English by EFL learners: Comparing the outcomes from paper and mobile application flashcards. *Journal of Computers in Education*, *9*(1), 93–111. https://doi.org/10.1007/s40692-021-00197-6

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental l2 vocabulary learning? A meta-analysis. *Language Learning*, *71*(2), 487–536. https://doi.org/10.1111/lang.12444

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. In *Psychological Bulletin* (Vol. 147, Issue 4). https://doi.org/10.1037/bul0000309

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 263–277. https://doi.org/10.1037/xap0000122

Yeşilbursa, A., & Bilican, R. (2013). Validation of self-regulatory capacity in vocabulary learning scale in Turkish. *Procedia - Social and Behavioral Sciences*, *70*, 882–886. https://doi.org/10.1016/j.sbspro.2013.01.134

Yüksel, H. G., Mercanoğlu, H. G., & Yılmaz, M. B. (2020). Digital flashcards vs. wordlists for learning technical vocabulary. *Computer Assisted Language Learning*, *0*(0), 1–17. https://doi.org/10.1080/09588221.2020.1854312

Zakian, M., Xodabande, I., Valizadeh, M., & Yousefvand, M. (2022). Out-of-the-classroom learning of English vocabulary by EFL learners: investigating the effectiveness of mobile assisted learning with digital flashcards. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(1). https://doi.org/10.1186/s40862-022-00143-8

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition*, *38*(8), 995–1008. https://doi.org/10.3758/MC.38.8.995

Za'rour, G. I., & Buckingham, T. (1969). Some factors affecting improvement in proficiency in English as a second language. *TESOL Quarterly*, *3*(1), 37. https://doi.org/10.2307/3586041

Zeiss, P. A. (1983). A comparison of the effects of super learning techniques on the learning of English as a second language. *Journal of the Society for Accelerative Learning and Teaching*, *9*(2), 93–101.

Zhang, B., & Li, C. (2011). Classification of L2 vocabulary learning strategies: Evidence from exploratory and confirmatory factor analyses. *RELC Journal*, *42*(2), 141–154. https://doi.org/10.1177/0033688211405180

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696–725. https://doi.org/10.1177/1362168820913998

Zhang, X., & Lu, X. (2014). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics*, *35*(3), 283–304. https://doi.org/10.1093/applin/amt014

Zhang, Y., Lin, C.-H., Zhang, D., & Choi, Y. (2017). Motivation, strategy, and English as a foreign language vocabulary learning: A structural equation modelling study. *British Journal of Educational Psychology*, *87*(1), 57–74. https://doi.org/10.1111/bjep.12135

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System*, *56*(220), 40–53. https://doi.org/10.1016/j.system.2015.11.007

Zou, D., Huang, Y., & Xie, H. (2021). Digital game-based vocabulary learning: where are we and where are we going? *Computer Assisted Language Learning*, *34*(5–6), 751–777. https://doi.org/10.1080/09588221.2019.1640745

APPENDICES

Appendix 1. The Updated Vocabulary Levels Test (Bilingual - Arabic)

اختبار قياس المفردات

الاسم الرباعي بالعربي: .....

الفصل: .....

يقيس هذه الاختبار معرفتك بالكلمات المفيدة في اللغة الإنجليزية. ضع علامة صح (✓) تحت الكلمة الإنجليزية التي تتناسب مع الكلمة العربية. انظر للمثال التالي:

	game	Island	Mouth	Movie	Song	Yard
جزيرة						
فم						
أغنية						

تكون الإجابة على الصندوق السابق بهذا الشكل:

	game	Island	Mouth	Movie	Song	Yard
جزيرة		✓				
فم			✓			
أغنية					✓	

1,000 Word Level

	choice	computer	garden	photograph	price	week
سعر						
صورة						
حديقة						

	eye	father	night	van	voice	year
عين						
أب						
ليل						

	center	note	state	tomorrow	uncle	winter
عم/خال						
وسط						
مذكرة						

	box	brother	horse	hour	house	plan
أخ						
ساعة						
خطة						

	animal	bath	crime	grass	law	shoulder
عشب						
حمّام						
كتف						

	drink	educate	forget	laugh	prepare	suit
يجهّز						
يضحك						
ينسى						

	check	fight	return	tell	work	write
يعمل						
يعود						
يتأكد						

	bring	can	reply	stare	understand	wish
يرد						
يحضر						
ينظر لشيء فترة طويلة						

	alone	bad	cold	green	loud	main
رئيسي						
سيئ						
بارد						

	awful	definite	exciting	general	mad	sweet
شيء محدد أو معروف						
عام						
شنيع						

2,000 Word Level

	coach	customer	feature	pie	vehicle	weed
ميزة						
مدرب						
عشب يابس						

	average	discipline	knowledge	pocket	trap	vegetable
خضار						
معرفة						
متوسط						

	circle	justice	knife	onion	partner	pension
دائري						
سكين						
عدل						

	cable	section	sheet	site	staff	tank
جزء						
موقع						
شرشف السرير						

	apartment	cap	envelope	lawyer	speed	union
ظرف						
قبعة						
شقة						

	argue	contribute	quit	seek	vote	wrap
يغلّف						
يساهم						
يطلب						

	avoid	contain	murder	search	switch	trade
يحتوي						
يبحث						
يتجنب						

	bump	complicate	include	organize	receive	warn
يستلم						
يصدم						
يشمل						

	available	constant	electrical	medical	proud	super
فخور						
رائع						
مستمر						

	environmental	junior	pure	rotten	smooth	wise
متعفن						
سلس						
مبتدئ						

3,000 Word Level

	angle	apology	behavior	bible	celebration	portion
سلوك						
احتفال						
اعتذار						

	anxiety	athlete	counsel	foundation	phrase	wealth
عبارة						
استشارة						
ثروة						

	agriculture	conference	frequency	liquid	regime	volunteer
زراعة						
نظام						
متطوع						

	asset	heritage	novel	poverty	prosecution	suburb
فقر						
موروث						
ذو قيمة ومفيد						

	audience	crystal	intelligence	outcome	pit	welfare
ذكاء						
حفرة						
جمهور						

	consent	enforce	exhibit	retain	specify	target
يوافق						
يحدد						
يعرض						

	accomplish	capture	debate	impose	proceed	prohibit
يقبض						
يتقدم						
يناقش						

	absorb	decline	exceed	link	nod	persist
يستمر						
يتجاوز						
يمتص						

	approximate	frequent	graphic	pale	prior	vital
تقريبي						
قبل						
متكرر						

	consistent	enthusiastic	former	logical	marginal	mutual
ثابت او متماسك						
سابق						
متبادل						

4,000 Word Level

	cave	scenario	sergeant	stitch	vitamin	wax
فيتامين						
كهف						
سيناريو						

	candle	diamond	gulf	salmon	soap	tutor
صابون						
معلم						
ألماس						

	agony	kilogram	orchestra	scrap	slot	soccer
فرقة موسيقية						
فتحة						
يقشر						

	crust	incidence	ram	senator	venue	verdict
قشرة						
حكم						
ساحة						

	alley	embassy	hardware	nutrition	threshold	tobacco
سفارة						
تبغ						
زقاق						

	fling	forbid	harvest	shrink	simulate	vibrate
يمنع						
ينكمش						
يرمي						

	activate	disclose	hug	intimidate	plunge	weep
يصيح						
يفشي سرا						
يفعّل البرنامج						

	diminish	exaggerate	explode	penetrate	transplant	verify
ينفجر						
يتقلص						
يزرع						

	adjacent	crude	fond	sane	spherical	swift
مجاور						
عاقل						
سريع						

	abnormal	bulky	credible	greasy	magnificent	optical
موثوق						
دهني أو مزيّت						
غير طبيعي						

5,000 Word Level

	gown	maid	mustache	paradise	pastry	vinegar
شنب						
فردوس						
معجنات						

	asthma	chord	jockey	monk	rectangle	vase
مزهرية						
نغمة						
مستطيل						

	batch	dentist	hum	lime	pork	scripture
ليمون أخضر						
يدندن						
لحم الخنزير						

	amnesty	claw	earthquake	perfume	sanctuary	wizard
عطر						
ساحر						
ملاذ						

	altitude	diversion	hemisphere	pirate	robe	socket
ارتفاع						
رداء						
قرصان						

	applaud	erase	jog	intrude	notify	wrestle
يخبر						
يقتحم						
يمسح						

	bribe	expire	immerse	meditate	persecute	shred
يقطع						
ينتهي						
يتأمل						

	commemorate	growl	ignite	pierce	renovate	swap
يشعل						
يبدل						
يثقب						

	bald	eternal	imperative	lavish	moist	tranquil
هادئ						
أصلع						
رطب						

	diesel	incidental	mandatory	prudent	superficial	tame
يروض						
اجباري						
حكيم						

Appendix 2. The Recall Updated Vocabulary Levels Test (Bilingual - Arabic)

اختبار قياس المفردات الإنجليزية الشائعة

الاسم الرباعي بالعربي: .....

الفصل: .....

يقيس هذه الاختبار معرفتك بالكلمات المفيدة في اللغة الإنجليزية. المطلوب منك ترجمة الكلمات الإنجليزية الى اللغة العربية كما في المثال التالي:

1. Good .....
2. Water .....
3. Apple .....

تكون الإجابة على الأسئلة السابقة كالتالي:

1. Good .....جيد.....
2. Water .....ماء.....
3. Apple .....تفاح.....

1,000 Word Level

1. price .....
2. photograph .....
3. garden .....
4. eye .....
5. father .....
6. night .....
7. uncle .....
8. center .....
9. note .....
10. brother .....
11. hour .....
12. plan .....
13. grass .....
14. bath .....
15. shoulder .....
16. prepare .....
17. laugh .....
18. forget .....
19. work .....
20. return .....
21. check .....
22. reply .....
23. bring .....
24. stare .....
25. main .....
26. bad .....
27. cold .....
28. definite .....
29. general .....
30. awful .....

2,000 Word Level

1. feature .....
2. coach .....
3. weed .....
4. vegetable .....
5. knowledge .....
6. average .....
7. circle .....
8. knife .....
9. justice .....
10. section .....
11. site .....
12. sheet .....
13. envelope .....
14. cap .....
15. apartment .....
16. wrap .....
17. contribute .....
18. seek .....
19. contain .....
20. search .....
21. avoid .....
22. receive .....
23. bump .....
24. include .....
25. proud .....
26. super .....
27. constant .....
28. rotten .....
29. smooth .....
30. junior .....

3,000 Word Level

1. behavior .....
2. celebration .....
3. apology .....
4. phrase .....
5. counsel .....
6. wealth .....
7. agriculture .....
8. regime .....
9. volunteer .....
10. poverty .....
11. heritage .....
12. asset .....
13. intelligence .....
14. pit .....
15. audience .....
16. consent .....
17. specify .....
18. exhibit .....
19. capture .....
20. proceed .....
21. debate .....
22. persist .....
23. exceed .....
24. absorb .....
25. approximate .....
26. prior .....
27. frequent .....
28. consistent .....
29. former .....
30. mutual .....

4,000 Word Level

1. vitamin .....
2. cave .....
3. scenario .....
4. soap .....
5. tutor .....
6. diamond .....
7. orchestra .....
8. slot .....
9. scrap .....
10. crust .....
11. verdict .....
12. venue .....
13. embassy .....
14. tobacco .....
15. alley .....
16. forbid .....
17. shrink .....
18. fling .....
19. weep .....
20. disclose .....
21. activate .....
22. explode .....
23. diminish .....
24. transplant .....
25. adjacent .....
26. sane .....
27. swift .....
28. credible .....
29. greasy .....
30. abnormal .....

5,000 Word Level

1. mustache .....
2. paradise .....
3. pastry .....
4. vase .....
5. chord .....
6. rectangle .....
7. lime .....
8. hum .....
9. pork .....
10. perfume .....
11. wizard .....
12. sanctuary .....
13. altitude .....
14. robe .....
15. pirate .....
16. notify .....
17. intrude .....
18. erase .....
19. shred .....
20. expire .....
21. mediate .....
22. ignite .....
23. swap .....
24. pierce .....
25. tranquil .....
26. bald .....
27. moist .....
28. tame .....
29. mandatory .....
30. prudent .....

Appendix 3. Self-Regulating Capacity in Vocabulary Learning' Scale (SRCvoc; Tseng et al., 2006)

Item	Learning experience
1.	Once the novelty of learning vocabulary is gone, I easily become impatient with it.
2.	When I feel stressed about vocabulary learning, I know how to reduce this stress.
3.	When I am studying vocabulary and the learning environment becomes unsuitable, I try to sort out the problem.
4.	When learning vocabulary, I have special techniques to achieve my learning goals.
5.	When learning vocabulary, I have special techniques to keep my concentration focused.
6.	I feel satisfied with the methods I use to reduce the stress of vocabulary learning.
7.	When learning vocabulary, I believe I can achieve my goals more quickly than expected.
8.	During the process of learning vocabulary, I feel satisfied with the ways I eliminate boredom.
9.	When learning vocabulary, I think my methods of controlling my concentration are effective.
10.	When learning vocabulary, I persist until I reach the goals that I make for myself.
11.	When it comes to learning vocabulary, I have my special techniques to prevent procrastination.
12.	When I feel stressed about vocabulary learning, I simply want to give up.
13.	I believe I can overcome all the difficulties related to achieving my vocabulary learning goals.

14.	When learning vocabulary, I know how to arrange the environment to make learning more efficient.
15.	When I feel stressed about my vocabulary learning, I cope with this problem immediately.
16.	When it comes to learning vocabulary, I think my methods of controlling procrastination are effective.
17.	When learning vocabulary, I am aware that the learning environment matters.
18.	During the process of learning vocabulary, I am confident that I can overcome any sense of boredom.
19.	When feeling bored with learning vocabulary, I know how to regulate my mood in order to invigorate the learning process.
20.	When I study vocabulary, I look for a good learning environment.

Note: Commitment control: items 4, 7, 10, 13; metacognative control: items 5, 9, 11, 16; satiation control: items 1, 8, 18, 19; emotion control: items 2, 6, 12, 15; environmental control: items 3, 14, 17, 20.

Appendix 4. Self-Determination Theory of Second Language Scale (SDT-L2; Alamer, 2021)

Why are you learning English?

Autonomous motivation

Intrinsic orientation

Because I enjoy learning English.

Because of the pleasure I get when hear and read English.

For the satisfaction I feel when I use English.

For the enjoyment I experience when I achieve a new goal in English learning.

Because learning English is a fun activity in and of itself.

Identified orientation

Because learning English is important for my personal growth.

Because learning English can open new opportunities and possibilities for me.

For the value it holds in my self-development.

Because learning English is important for my current and future studies.

Because learning English allows me to read and hear English-based materials that are necessary for my personal success.

Controlled motivation

Introjected orientation

Because I would feel guilty if I didn't understand English.

Because I would feel ashamed if I'm not successful in English learning like my friend(s)/family. Because people around me (the teacher/peers/parents) expect me to learn English.

Because people around me (the teacher/peers/parents) would think I'm a failure if I didn't speak English.

Because I feel pressured by the people around me (the teacher/peers/parents) to learn English.

External orientation

Because I want to get a prestigious job that requires English proficiency.

Because I want to get better marks in the English course.

Because English is just a required course that I want to pass.

Because I don’t want to fail the final exam in the English course.

Because there will be negative consequences if I fail to learn English.

Appendix 5. Frequent quizzes and app vocabulary learning questionnaire.

Perceived app effectiveness scale

Vocabulary learning from the app was very useful.

I learned many words from the app.

Having the words in my phone made it easy to learn them.

App enjoyment scale

Vocabulary learning from the app was fun.

The app design promotes vocabulary learning.

The vocabulary learning experience from the app was joyful.

Perceived quizzes effectiveness scale

Quizzes helped me learn vocabulary.

Having quizzes motivated me to learn vocabulary.

Quizzes enhanced my vocabulary learning.

Quizzes enjoyment scale

Quizzes were annoying.

I enjoyed quizzes during the semester.

I was glad that we had quizzes.

Future app learning

I will keep learning vocabulary from the app in the future.

I have no intention to stop learning vocabulary from the app.

Vocabulary learning from the app will become part of my future routine.

Appendix 6. Linear models output for growth analysis

Recognition growth linear model

Fixed effects	β	Std. Error	Z value	p
(Intercept)	14.39	3.78	3.81	**< .001**
biweekly	-0.46	5.60	-0.08	0.935
monthly	0.54	5.71	0.09	0.925
no-quiz	-14.78	5.04	-2.93	**0.005**

Recall growth linear model

Fixed effects	β	Std. Error	Z value	p
(Intercept)	4.65	2.78	1.67	0.098
biweekly	4.98	3.98	1.25	0.215
monthly	-1.58	4.33	-0.36	0.716
no-quiz	-3.74	3.80	-0.98	0.328