**

The acquisition of novel word meanings from recreational reading under massed and distributed learning conditions

-

A dissertation submitted by

Colin Christopher MacCallum Wray, B.A., M.A., LL.B.

In partial fulfillment of the requirements for the award of

Doctor of Education

Faculty of Education, University of Southern Queensland

2014

Certification of Dissertation

I certify that the ideas, experimental work, results, analyses, and conclusions reported in this dissertation are entirely my own effort, except where otherwise acknowledged. I also certify that the work is original and has not been previously submitted for any other award, except where otherwise acknowledged.

Signature of Candidate Date

ENDORSEMENT

Signature of Supervisor/s Date

Acknowledgements

I would like to thank Dr. Shirley O’Neill and Dr. Nicole Green as well as Professor Patrick Danaher, all of whom have offered encouragement and advice at various stages during the writing of this dissertation. Their support was very much appreciated and useful in bringing to my attention various errors that had eluded me. It was a constant reassurance to know that academics of this caliber were overseeing the dissertation project, and it would not have been possible to see this study through to completion without their timely interventions and suggestions. I can’t help but view the dissertation as a team effort: Yes, the ideas were mine, and responsibility for any errors rests with me alone, but without such valuable assistance and suggestions at timely junctures, I wonder whether I would ever have satisfied my external examiners.

My family has been supportive throughout the whole project, and I am grateful for their patience in letting me get on with typing out chapters when I should have been completing any number of other tasks around the house and elsewhere. My mother, has been supportive and helpful -full of valuable advice and never failing to point out an error. Thanks for keeping me on track.

A special thanks to three teaching assistants (who wish to remain anonymous) who played an important role in helping to conduct tests and compile the data from which the current study derives its conclusions; their professionalism and dedication contributed significantly to the study’s completion. I also owe a great deal to a colleague, wishing to remain anonymous, who helped immeasurably during the process of data gathering.

Contents

Chapter 5: Results and findings

List of Appendices

List of Tables

List of Figures

Abstract

Children learning English as a second language acquire much vocabulary from recreational reading (RR) as do their English native-speaking peers. Such learning typically involves a cycle of repeated encounters with the same novel word in different contextual settings, each encounter consolidating and building upon prior knowledge (Nation, 1990). This dissertation examines one factor that potentially impacts upon the pedagogical value of RR as a vocabulary-building practice: the time intervals between the reader's encounters with the same novel word while engaged in in-class RR sessions. The study makes use of five sets of texts, each designed to expose the reader to a uniquely more, or less, distributed encounter with a small sample of non-words particular to the set in which they occur.

Employing a researcher-designed data-elicitation instrument (the Vocabulary State Assignment Task (VSAT)), the study demonstrates that among a population of Thai primary school English as a second language (EAL) children, distributed encounters with novel non-words potentially lead to more impressive meaning gains of those same non-words than do massed encounters (i.e., many encounters with the same word over a relatively short time period). Drawing upon three alternative (reasonable) notions of what it means to know a word, the investigation demonstrates that each implies different learning outcomes in terms of (a) whether total words gained differed significantly from having read one set of texts as opposed to another and (b) how substantial were those differences should they arise.

A breakdown of target word gains by lexical class (nouns, verbs, adjectives, and adverbs) revealed an association between how distributed were occurrences of the same novel word and statistically significant differences in the proportions of learned words of the particular class of interest. The study finds that while spaced presentations could account for disparities in the sums of nouns and verbs children 'knew' to a particular notion of known

($p<0.05$), this does not hold true for either adjectives and adverbs. Whether differences in noun and/or verb totals (and, by implication, learning) proved significantly different from encountering the same word under more, or less, spaced conditions, and how substantial were any such differences, depended upon the definition of known one acknowledges.

Chapter 1

Introduction

Preface

The origins of this research lie in a longstanding professional interest in vocabulary instruction dating back to the late 1980s when I first set out on the path as an English language teacher. In those 'far off' days, linguists viewed second language (L2) mastery as a subconscious process that drew upon an innate ability to derive language competence from comprehensible input (Krashen,1987). Rather than direct instruction, teaching would center around supplying examples of well-formed utterances related to themes from which acquisition could proceed. Language classes positively exuded informality and an air of student centeredness. In one lesson, the classroom might 'stand in' for a railway station, students randomly mingling and engaging in short exchanges to buy tickets; in the next, a busy street, with learners asking for and giving directions. Explicit teaching had little place in this humanistic melee –indeed, many of my more experienced colleagues would, and did, take every opportunity to vocally disavow it. But if the typical class would surely have come across as distressingly unstructured and unpredictable by today's standards, linguistic theory left little doubt as to how language mastery unfolded: Students acquired word meanings naturalistically from language exposure just as they did other aspects of language –grammar, pronunciation, intonation, and so on.

By the late 1980s, with critics of this, 'the communicative approach' becoming more vocal, and arguments for explicit instruction more persuasive, vocabulary drills and exercises once again began to make their way into English as an Additional Language (EAL) classrooms. At my own school, student needs foremost in mind, we cautiously introduced a structured vocabulary teaching program based upon Willis's (1990) lexical syllabus and began actively prioritizing vocabulary study. The shift towards formal instruction soon proved effective, with test scores and anecdotal evidence revealing impressive and durable word knowledge gains.

Staff enthusiasm for our pedagogical re-focus did not come without caveats, however. Teaching word meanings proved time consuming which meant cutting down on popular communicative activities. To some, lessons had lost something of their spontaneity, naturalness and, dare I say it, fun! Gone was the fluidity, jollity and innocent humor of everyday communicative interaction. In its place came complex technical discussions of shallow nuances of meaning. As teachers, we could see that explicit vocabulary instruction 'worked.' Yes, it yielded tangible and worthwhile gains. Even so, we wondered whether more humanistic and child centered forms of instruction might prove equally effective. The suggestion, floated from time to time, that recreational reading[1] *(RR) –i.e. pleasure reading- might serve as a powerful source of vocabulary gain hardly amounted to a novel or radical proposition, McCracken (1971) having argued as much since the early 1970s. Like others, however, I questioned whether such a familiar old staple could play other than a supporting role to our formal teaching efforts.*

Several years on, and by now working in one of Thailand's new international schools, an opportunity to evaluate the effects of reading first hand presented itself when we trialed RR in our primary division –partly to address vocabulary concerns but also raise literacy skills more generally. From the outset, children clearly enjoyed their reading opportunities and teacher feedback proved highly favorable. When it came to reviewing our new initiative some several months in, we agreed that RR sessions should continue; rarely, indeed, had teachers shown such wholehearted support for a timetable revision. At the same time, no one seriously denied that important issues lay unresolved. Was recreational reading quite as productive as it could be? What scope existed to develop this practice into something more powerful through textual adaptation, and what role might explicit instruction play as a supplement to RR sessions? For myself, a single question intrigued me above all: If children indeed gained new words from pleasure reading, then upon what factors might the breadth and depth

[1] Explicit instruction because it affords opportunities for readers to apply strategies for word-meaning derivation.

of such gains depend? In the light of the well-documented spacing effect (see e.g. Dempster, 1988) I wondered whether the time intervals between encounters with the same novel word might significantly impact learning outcomes. On this critical question, however, the literature remained curiously silent.

1.1 Introduction

During the last 10 or so years, popular publications (e.g., Krashen, 2004; Pilgreen, 2000) along with findings from a sizeable body of research (e.g., Garan & DeVoogd, 2008; Kirby, 2003; Trelease, 2006) have persuaded several international schools in Thailand to set aside a fixed period during the school day for children to read pleasurable and challenging texts of their choice. These reading opportunities come under the terms *Extensive Reading*, *Sustained Silent Reading*, *Uninterrupted Sustained Silent Reading* (USSR), *Drop Everything And Read* (DEAR) or, more generically, *Recreational Reading* (RR). Literacy gains from RR have proved broad ranging. They include improvements in writing (Janopoulous, 1986; Mason, 2007), vocabulary (Cho, Park, & Krashen, 2008; Saragi, Nation, & Meister, 1978), reading fluency (Readence, Bean, & Baldwin, 2004), comprehension (Elley, 1991; Smith, 1994), spelling (Hafiz & Tudor, 1990; Nisbet, 1941), and grammatical awareness (Rodrigo, 2006; Stokes, Krashen, & Kartchner, 1998). Such benefits, moreover, typically arise irrespective of the reader's age, language background or grade level. RR has proved effective in both primary and secondary school settings, and among English native speakers as well as those learning English as a second language (Cho & Kim, 2004; Krashen, 2004; Liu, 2007).

Although dating to the 1970s (Chua, 2008; Krashen, 2004), few studies have addressed optimizing RR for literacy gain, thereby leaving important pedagogical questions unanswered: For which populations might RR prove most beneficial? Will an RR program prove cost effective? Does RR raise certain language skills more so than others? And to what degree should RR supplement more formal literacy instruction? (see VanDeweghe, 2008). These issues continue to attract lively staffroom debate (see, e.g. Garan, 2008) as has the more general question concerning the proper place of RR in language arts programs and its efficacy as a pedagogical practice (Klump, 2007; Krashen, 2004).

The current dissertation explores a narrow, yet relevant, topic of interest to those exploring the potential of recreational reading as a teaching practice –the contribution of RR to reader uptake and consolidation of novel content word meanings in written texts. Motivated by findings from spacing effect research (Dempster, 1988), the dissertation examines the impact on the sum of children's word meaning gains of the time intervals between reencounters with the same novel word during RR sessions. Drawing upon the RR experiences of volunteer Thai (L1) primary school EAL learners and sets of adapted texts, the dissertation asks whether these intervals explain significant differences in children's acquisition of new word meanings from their RR sessions. The investigation sets out to quantify any such observed differences and evaluate their statistical and pedagogical importance.

1.2 The spacing effect

In the late 1890s, the psychologist Hermann Ebbinghaus demonstrated that interspersing learning opportunities with short 'breaks' or 'distractions' ensured substantial time savings in tasks that involved learning semi-random letter strings to a criterion standard. He termed this finding the 'spacing effect,' a label that captures the notion of a 'gap' or 'interval' between successive learning opportunities. Ebbinghaus' later works explored the *time savings* and the *durability* of learning outcomes under spaced learning conditions. The first line of inquiry supplied estimates of reduced study time arising from rearranging the timing and duration of intervals, while the second yielded the 'forgetting curve,' a representation of time-associated memory decay following a successful learning effort. Today, some 125 years since these early pioneering studies, the spacing effect has proven both pervasive and robust. The effect arises with non-lexical tasks (paragraph recall, motor-skill acquisition, mirror tracing, etc.[2]), from inductive as well as deductive learning opportunities, and among participants of various ages, aptitudes, and educational experience (Vlach & Sandhoffer, 2012). The

[2] See Thalheimer (2006) for a broad-ranging review of the literature.

time savings from spaced learning remain one of the most readily demonstrated findings of cognitive psychology (Dempster, 1988).

Yet despite its obvious applications to vocabulary teaching (see e.g. Bloom & Shuell, 1981) the evidence for a spaced learning advantage still derives primarily from laboratory investigations as opposed to classroom based research. In the typical ‘clinical’ study, participants receive a list of word/meaning pairs to learn within a specified time period. At predetermined points the researcher interrupts the memorization effort by introducing a non-study interval (NSI) –a period of anything from a few minutes to several hours– after which the learning task resumes. The study concludes with a memory test requiring participants to recall as many words as possible from the listed items. From the ‘retained’ words, the researcher now attempts to relate learning outcomes to the frequency and duration of the timed interruptions (i.e. NSIs), along with total expended study time. Follow up investigations might explore changes in the number of non-study intervals, alternative interval durations, or longer/shorter time spans between NSI placements. School-based studies, despite their high ecological validity, remain uncommon. To date, no research has examined the potential to maximize vocabulary gains from RR through controlling when children reencounter the same target word during reading sessions.

1.3 Research interest and significance of the dissertation

The study addresses an important, pedagogical and theoretical issue:

Will primary-aged EAL children engaged in a school-based RR program learn significantly more vocabulary from frequent (or massed) textual reencounters with a novel word of interest (i.e., many encounters during a single RR session) than from encounters dispersed over a broader time span (e.g., RR sessions distributed over several days)?

The issue appears in diagrammatic form in Figure 1 (below) that depicts (1.) massed, (2.) less massed, and (3.) distributed target word (“x”) occurrences in a hypothetical text. In the massed condition (Case 1), the reader encounters 12 instances of a target word in a single reading session. In the distributed condition, those 12 encounters occur over six sessions, and in the less massed condition, over three. The dissertation asks whether more, or less, massed reencounters with target words during recreational reading result in meaningful differences -both statistically and pedagogically- in the number of words children learn assuming alternative, plausible, notions of what ‘word knowing’ might reasonably imply.

Massed learning

Session 1
XXXXXXXXXXXX

Less massed learning

Session 1	Session 2	Session 3
XXXX	XXXX	XXXX

Distributed learning

Session 1	Session 2	Session 3	Session 4	Session 5	Session 6
XX	XX	XX	XX	XX	XX

Figure 1: Distributed and massed learning.

The study’s pedagogical relevance stems from the potential teachers enjoy to exercise control over the intervals between word reencounters. This control comes from decisions regarding when reading occurs (on what days per week, and how many times per day), the duration of reading sessions, or from textual adaptations affecting the time that elapses before a student meets the same novel word during a reading opportunity. Such adaptation might include, for example, inserting or deleting words/clauses or whole paragraphs from scripts (see Chapter 4). For those interested in the theoretical task of model building, the research asks whether spaced learning merits inclusion in a comprehensive and explanatory account of vocabulary gain from reading experiences.

Apart from its primary goal of identifying and quantifying the effects of the time intervals between word reencounters, the study also contributes to topical discussion on a range of teacher concerns. The dissertation describes the challenges of adapting texts to maximize the potential of spaced learning opportunities; it affirms incidental word learning among Thai EAL children from RR sessions (indeed, it stands as the first to do so), and it presents a robust methodology for exploring spaced learning with tertiary level students along with gifted children and those on the special needs register, neither population having participated in the current research for reasons noted in chapter 4. Not least, the study contributes towards furthering our understanding of RR as a pedagogical practice. Hopefully it will promote a more nuanced/informed discussion of RR's proper place in the primary school curriculum (see Garan & DeVoogd, 2008 for a discussion).

1.4 What it means to know a word and how to measure word knowledge

In *Learning Vocabulary in Another Language*, Nation (2001) cautions against regarding words as decontextualized, "isolated," units of language (p.23). He reminds us that a word typically 'associates' with others, occurs in particular contexts (e.g. 'merry' with Christmas), appears in certain grammatical constructions (the verb 'happen' only appears in active voice sentences), and preferentially occurs in certain registers (e.g. "Hi!" in informal as opposed to formal speech). What it means to know a word remains a complex issue that demands answers to contentious and non-trivial questions (Nation, 2001; Waring, 1999): How many categories of lexical understanding, for example, should we acknowledge? What significance should attach to each? (Should we regard knowing how to spell a word as more important than its correct pronunciation?) and how ought we to discriminate between permissible and aberrant examples of word usage? Views on such issues continue to evolve in the light of new research findings. In the 1940s most linguists associated *word knowing* with a handful or so of basic understandings. Cronbach (1942), for example, cited just 5 knowledge types: (1.) *Application knowledge* –the capacity to select an appropriate use of the word, e.g. "food" as opposed to "cuisine" during an informal discussion; (2.) *Breadth knowledge* –the term refers to the language user's familiarity with a word's different meanings and usage, e.g. that "license" functions either as a verb or noun; (3.) *Precision knowledge* –the ability to select a word most appropriate for communicating the intended message; (4.) *Availability knowledge* –the ability to use the word productively, and (5.) *Generalization knowledge* –i.e. familiarity with the word's primary, or standard, meaning. By the mid-1970s, and building upon Cronbach's (1942) work, Richards (1976) proposed his influential eight key assumptions, a pedagogically centered characterization of word knowledge that would "help … determine the status of vocabulary teaching within the syllabus" (p.78). To Cronbach's (1942) listing, Richards (1976) added awareness of the word's frequency of occurrence, its register and the syntactic structures in which it occurs. These assumptions, in turn, informed an altogether more extensive and comprehensive conception of *word knowing* from Paul Nation (2001).

Nation (2001) categorizes word knowledge into two basic types, receptive and productive,[3] a distinction Cronbach (1942) and Richards (1976) had largely neglected, then goes on to subdivide each category in terms of *meaning* (e.g., what the word refers to), *form* (whether it occurs in spoken or written language), and *use* (e.g., the communicative circumstances in which it typically arises).[4] The productive competencies relate to the 'so-called' active skills of speaking and writing skills and include, among others, the ability to spell, identify collocations, pronounce words correctly, and select appropriate words from among alternatives given the context and setting of the communication. The receptive competencies, conversely, refer to the 'passive' skills of reading and listening. They include familiarity with the word's phonology (i.e., what it 'sounds' like), its orthography (constituent letters), the concept(s) it denotes, its idiomatic extensions, and the language user's awareness of the grammatical constructions it most commonly associates with.[5]

With a few exceptions such as 'part of speech,' linguists view lexical competencies as points lying along their own particular measurement scales. A child might understand something, for

[3] Broadly, receptive skills are those that serve the purposes of reading and listening, while productive skills allow for writing and speaking.
[4] The grammatical patterns in which it occurs; the words with which it commonly appears (collocations); and where, when, and how often we expect to meet the word (Coxhead, 2010).
[5] Note that the same competence may be productive and receptive. Knowing a word's meaning is an obvious example.

example, of a word's 'meaning' but fall short of grasping its full meaning specification, or correctly pronounce a word's base form but not its inflectional variants. Proficiency in one aspect of word knowledge may not indicate mastery of another. Those with excellent pronunciation may have minimal awareness of collocations, just as capable spellers might remain ignorant of the contexts in which a word correctly arises. To derive an overall measure of a language user's lexical competence calls for evaluating both how many words that language user objectively 'knows' (i.e. the issue of 'knowledge breadth') and the extensiveness of that understanding (i.e. 'knowledge depth'). Breadth measurement typically involves administering multiple-choice tests (MCTs) of random words from various frequency bands of occurrence such as, for example, the first thousand most common words, the second thousand most common, and so on. Nation's (1983) Levels tests of productive and receptive vocabulary employ this assessment methodology as does the Peabody Picture Vocabulary test (2009). Depth measurement, conversely, tends to draw upon word association tests (see, e.g., Read, 1993, 2004 on the design of this test type) or one-to-one discussions in which the researcher assesses understanding from student responses to guided questions.

More elaborate measurement approaches expand upon depth and breadth to incorporate additional dimensions of word understanding, or more advanced notions of what 'breadth' and 'depth' reasonably imply. In Henriksen's (1999) multi-continua model, for example, a 'word-knowledge' category captures a language user's familiarity with word meaning, while a second category, 'use-knowledge' describes how efficient is word retrieval from memory. A third category, 'knowledge depth,' denotes familiarity with paradigmatic (antonymy, synonymy, hyponymy), syntagmatic (collocational restrictions) and other sense relations (Gao, 2013). Henriksen (1999) approaches 'knowledge' quantification by assigning each knowledge type its own measurement scale i.e. proficiency continuum. For 'word knowledge,' he proposes the 'partial precise' continuum along which lie degrees of denotational understanding. A 'use continuum' delimits the range of possible values for 'use knowledge,' while a third continuum, 'depth of knowledge,' denotes degree of 'vocabulary depth' i.e. a language user's familiarity with the associations between one word and others. Where points lie along each continua at the time of interest defines the language user's current lexical understanding (Henriksen, 1999).

For teachers and linguists, multi-continua models allow reporting and analyses of vocabulary knowledge within (1990) multi-dimensional conceptions of 'word knowing' that recognize discrete aspects of what 'to know' a word reasonably signifies. The models helpfully reveal in which competence a learner exhibits more proficiency than another, the knowledge types upon which to focus instruction, and how lexical development is proceeding. A learner's rate of progress along a continuum typically depends primarily upon three factors: (1.) textual features of the reading materials s/he engages with, (2.) the student's affective disposition, and (3.) the cognitive factors and topic knowledge the learner brings to the reading task (Rydland, 2012). The textual features consist of structural properties of scripts that impact upon its comprehensibility. Examples include the informativeness of contextual clues (Ma, 2008; Rankin & Overholser, 1969), the proportion of unknown lexis (Laufer, 1992), average word length (Nagy, Anderson, & Herman, 1987), genre (Shokouhi & Maniati, 2009), the frequency of novel word recurrence (Pigada & Schmitt, 2006), concreteness (Paivio, 1991; Salsbury, Crossley, & McNamara, 2011), and incidentals such as font size and letter shape (Tavakoli & Kheirzadeh, 2011). The term 'affective disposition,' refers to the reader's emotional state while engaging with the text –whether s/he finds the text motivating (Krashen, 2004; Swanborn & deGlopper, 1999; Williams, 1994), anxiety raising (Sellers, 2000), or pleasure inducing (Pilgreen, 2000). In general, RR will most effectively raise vocabulary when readers feel relaxed, attentive, and find themselves engaged with enjoyable texts –a view Krashen (1988) expresses in his 'pleasure hypothesis.' The third class of factors, the 'cognitive,' comprises the mental attributes readers draw upon to match orthographic strings with phonological representations and then integrate the meanings of those representations into their understanding of the script. These attributes include working memory capacity (Papagno, Valentine, & Baddeley, 1991), reading ability (Rinehart, Stahl,

& Erickson, 1986), verbal IQ (Smith, Smith, Taylor, & Hobby, 2005), and aptitude for selecting and applying word-meaning derivation/preservation strategies (Nemati, 2009; Siriwan, 2007). The task of synthesizing these various text, affective, and cognitive factors into a viable explanatory model of vocabulary learning from reading experiences remains a popular research topic.

1.5 Methodology and limitations

This study focuses on a single but important aspect of word knowledge –word meaning.[6] Following Bloom (2000), a child demonstrates familiarity with a meaning once s/he has objectively established a match in long-term memory between a letter string, or phonological form, and the concept to which that string or form refers (p.1101). A child 'knows' the word [book], in this view, when s/he understands the properties books share with one another, along with features that disambiguate this reading material from other scripts such as, for example, magazines, newssheets, pamphlets, etc. Should a child report that [book] might serve as a verb as well as a noun, or that the term extends to electronic publications (eBooks), then s/he displays a deeper knowledge than another without such awareness (Qian, 1999).

The study addresses the Research Question (Section 1.3) by tracing young (9-year-old) Thai first language (L1) EAL primary students' (n=28) acquisition and retention of 20 non-words embedded in five sets of specially prepared booklets each constructed to expose a reader to 12 instances of four non-words from the aforementioned 20 (one noun, verb, adverb, and adjective, specific to the set in which they occur). A specially designed timetable specifies when reading takes place and ensures a set of texts requires a unique number of daily (35-minute) RR sessions for completion. Depending upon the set, a child encounters the 12 repetitions of each target word under more, or less, spaced learning conditions than from reading another. These conditions range from a highly massed that

[6] More particularly, it is word denotational gains from the process of assigning meanings of L1 words to new, and unfamiliar, L2 orthographic or phonological forms.

exposes the reader to all target word repetitions (i.e. 12 repetitions of each of the four words embedded in that set of texts) during a single RR session (Set 1), to highly distributed in which the 12 reencounters occur over the course of five consecutive daily sessions (Set 5). The remaining sets (Sets 2, 3, and 4) provide intermediate degrees of spaced learning, exposing the reader to the 12 instances of each target word over time periods ranging from between two (Set 2) and four (Set 4) consecutive days.

To assess a learner's word knowledge gains, the study employs the Vocabulary State Assignment Task (VSAT), an instrument that allocates each target non-word to one of six knowledge states. Each state represents a step in a hierarchical scale ranging from total unfamiliarity with the target word (State 1) to 'high level' receptive and productive competence (State 6). Data from VSAT tests undertaken promptly after completing a set of texts supplied the raw statistics employed to answer the Research Questions (see Chapter 4). A full description of the instrument, the manner of its design, and the results of reliability tests appear in Chapter 3.

Findings (Chapter 5) from the present study apply only to the population from which data derives, along with any larger population of which participants might reasonably represent an unbiased sample (namely, the 'average,' Year 4, Thai EAL student –see Chapter 4). The dissertation does not claim relevance to native English L1 speaking pupils or those attending secondary education. Nor does the study necessarily offer insights relevant to children disadvantaged by low socioeconomic status or receiving support for Special Educational Needs (SEN). Chapter 4 (methodology) justifies and clarifies the selection of the student populations involved in the present research.

1.6 Terms and definitions

The following terms appear in the dissertation:

Acquisition: Acquisition refers to the "appropriation of information without awareness on the part of the acquirer of what is acquired and stored in implicit memory" (Paradis, 2009, p.4).

Contextual clues: This refers to objective indications of word meaning embedded in the text. Examples include affixes, verbal inflexions, pictures accompanying a script, and syntactic markers signifying word class along with general information imparted by the textual message itself; the definition excludes understandings the reader brings to the reading task through subject knowledge familiarity.

Declarative memory: This stands as one of the two human memory stores (the other being the procedural). The declarative memory store holds memories that language users consciously recall and that generally derive from conscious learning efforts (e.g., knowledge of various facts and events).

EAL (English as an Additional Language): The term refers to a learning context in which a child or adult attempts to gain proficiency in English having already mastered his/her L1.

Explicit learning: Following Jackson and Jackson (1995), this denotes "learning (as a product) which arises from conscious attempts to construct a representation of the task." Such learning stems from a "directed search of memory for similar or analogous task relevant information," and a conscious attempt "to derive and test hypotheses related to the structure of the task" (p.2).

Procedural memory: This constitutes the second type of human memory, the other being declarative. It comprises a store of subconsciously gained skills and competencies acquired without conscious attention to the stimuli from which those gains arose.

Implicit learning: The term refers to learning that arises without awareness of the stimuli from which that same learning derives. This form of learning consists of "task relevant information … acquired automatically and without conscious awareness of what is being learned" (Jackson & Jackson, 1995, p.2).

Incidental learning: In the context of vocabulary acquisition the term refers to any learning that arises from readers focusing upon deriving information from a text as opposed to a deliberate effort to learn, consolidate, or develop an aspect of vocabulary knowledge as such.

L1: A language user's first language (i.e., the language acquired before all others).

L2: The second language an adult or child masters.

Non-words: These comprise words that conform to English language phonotactic and morphological rules though do not appear in unabridged English language dictionaries; the term non-word contrasts with 'nonsense words,' the latter denoting unreal concepts: triangles with four sides, green flamingos, etc.).

Reading volume: The number of words read per period of time (week, day, year, or month).

Recreational reading (RR): In the current study, RR refers to pleasure reading of texts a child selects (or a teacher selects on the child's behalf) and which lie within an optimal reading difficulty. Such reading does not involve overt post-reading assessment or other forms of student accountability.

Word: The study takes the terms word, and word family, as synonyms. A word, therefore, consists of a base form (e.g. happy) along with its permissible inflectional variants (happier, happiest etc.).

1.7 Outline of the dissertation

Chapter 2 reviews key research relevant to how children gain word knowledge from reading experiences.[7] After introducing key constructs (attention, consciousness, and *noticing* in particular), the chapter defines the forms of learning operative during RR sessions and specifies their unique contributions to developing receptive and/or productive language skills. The discussion then turns to report upon the potential time savings under spaced learning conditions and explore the several factors (e.g., chronological age, L1 background, and depth of stimuli processing) that potentially moderate the depth and breadth of such gains. The chapter concludes by integrating findings into a comprehensive account of word-meaning gains from RR sessions.

Chapter 3 (**Methodology 1**) delves into the challenges of measuring and tracking word meaning gains within the context of classroom-based research. The discussion begins by evaluating a well-known data-elicitation instrument, the Vocabulary Knowledge Scale (VKS) (Paribakht &

[7] Note: The absence of prior studies exploring spaced learning effects upon RR precludes a detailed review of previous research. The focus of the dissertation on what amounts to a novel research topic means the chapter primarily concerns itself with establishing pointers to inform the design of a robust methodology, and locating the interest of the study within the context of how word gains arise.

Wesche, 1993) for use in the present study. The chapter reports upon several non-trivial concerns relating to VKS usage before going on to argue that a variant of this instrument, the VSAT (designed by the current researcher), should supply robust data for addressing the Research Questions. The chapter concludes by presenting the results of three small-scale investigations that aim to validate the VSAT for classroom-based research with child participants.

Chapter 4 (**Methodology 2**) begins by presenting two Research Questions (derivatives of the overarching Research question –see Section 1.3), and associated hypotheses, that form the focus of the current investigation. The discussion attempts to justify and explain key features of the experimental design and manner of execution. Topics discussed include: (1.) details regarding participant selection, (2.) the properties of the experimental texts, (3.) the design and integration of non-words into the reading materials, (4.) completion dates for various reading and test sessions, and (5.) a description of the statistical approaches to data collection and analysis (principally, Friedman's ANOVA, Cochran's Q, McNemar's Test, and the Sign Test).

Chapter 5 (**Results and Findings**) consists of three parts. Part 1 presents the findings of the investigation, drawing upon simple graphical displays and short accompanying explanatory texts. The chapter reports the statistical significance of test determinations, along with any incidentally arising points relevant to the Research Questions. The discussion, in particular, identifies those particular pairs of sets from which participants gained statistically (significantly) more words –either overall or of a lexical class– from reading one set of texts as opposed to another. Part 2 goes on to examine these 'set pairs' in more detail, moving beyond known word sums to establish what such differences in learning outcomes represent in terms of lexical competencies: whether gains amount primarily to productive or receptive skills, for example, and in what ratio. Part 3 consists of a speculative attempt to provide insights into pedagogical implications. This exercise involves estimating the 'per-academic-year' difference in (1.) gross sums of known words, and (2.) known words differentiated by lexical class, assuming children were to have read materials designed to the specifications of texts the study employs as opposed to their ‘regular’ reading materials. The one year period recommended itself as traditionally defining the ‘start’ and ‘end’ dates of educational programs.

Chapter 6 (**Implications and conclusion**) examines the findings of this dissertation within the broader context of well-established beliefs that have historically, and continue, to inform current understanding of the vocabulary learning process and teaching practices. A closing section moves on to explore the applicability of the present study to the circumstances under which the host institution operates. The section also touches briefly upon the relevance of findings to how Thai international schools in general might attempt to resolve vocabulary deficits among their EAL cohorts.

1.8 Summary

The chapter began with a brief summary of literacy gains from RR experiences (1.1), noting the paucity of research into optimizing RR for vocabulary development. One issue in particular that has, as yet, attracted little research concerns the possible relationship between the time interval between a reader’s reencounter with the same novel word, and the likelihood of gaining its meaning. The chapter argued that such intervals should reasonably excite theoretical and practical interest: theoretical, because of an obvious relevance to developing a comprehensive model of word gains from RR, and practical in that teachers exercise control over such intervals’ frequency and duration through textual adaptation and/or the timing of reading sessions. The chapter went on to briefly outline a methodology to explore the effects of spaced word presentations on learning outcomes.

Chapter 2

Vocabulary gain via reading

2.1 Introduction

This chapter begins with a broad-ranging review of theoretical and empirical findings that address, or touch upon, how children gain vocabulary from RR experiences. It locates the interest of the present investigation within the broader category of factors that affect the word-learning process, it identifies factors moderating the impact of spaced word presentations on learning outcomes, and draws attention to the limited research into spaced learning within the context of RR sessions and vocabulary development. Building upon this review, the chapter then goes on to construct a comprehensive account of vocabulary gain from RR experiences with reference to spaced learning effects on mastering word-meanings. Findings from the chapter will play a critical role in informing the design of a robust methodology to address the research issue of interest (Section 1.3).[8]

2.2 The mental processes involved in vocabulary learning

Central to understanding how children learn are the twin constructs of 'selective attention' and 'consciousness.' Selective attention refers to the behavior children (and adults) engage in to separate out useful stimuli for processing from competing background distractions and irrelevancies (Sarter & Lustig, 2009) –it is the behavior that allows a school child to focus on a text, for example, and disregard the hum of the overhead fan, or ignore the chatter of students in the corridor. As James (1890) explained some 120+ years ago, selective attention serves a filtering function that enables a child/adult to take "possession by the mind in clear and vivid form," of one out of several "simultaneously possible objects or trains of thought" (pp.403-404). More recent research (see e.g. Posner, Walker, Friedrich &

[8] For the specific Research Questions, the reader is referred to Section 4.2.1.

Rafal, 1987; Posner, 2012) has identified discrete mental states that underlie this focusing/filtering effort. These include: *alertness* in the sense of readiness to deal with whatever the senses perceive, *detection* (i.e., defined as the ability to cognitively register stimuli) and *orientation* (i.e., the ability to align attention to certain stimuli as opposed to others). *Orientation* itself divides into three sub-behaviors: (1.) disengaging from a stimulus (e.g. looking away from a letter string while reading), (2.) shifting to a new one (focusing on a new string), and (3.) Re-engaging with the new stimulus (Posner and Rothbart, 1992; Velmans, 1991).

By the 1980s cognitive psychologists (see e.g. Jackendoff, 1987) were beginning to distinguish between 'attention' as James (1890) had conceived of it, and 'consciousness,' the latter equated with awareness of stimuli as a subjective experience: i.e. how we 'feel'(happy, sad, angry etc.) in regard to stimuli the sensory system alerts us to. Despite the difficulties with disambiguation (Schmidt, 1995, p.18), and suggestions the terms may stand as synonyms (Carr & Curran, 1994), psychologists today acknowledge attention and consciousness as qualitatively different attributes. Attention refers to the purely 'computational' aspect of mind i.e. the mind conceived of as a mechanistic computing facility or machine (Schmidt, 1995). A pocket calculator attends to 'input' from this perspective, as does a fire alarm or a stopwatch. Consciousness, or its common synonym 'awareness,' conversely, denotes what Jackendoff (1987) called the phenomenological mind of sensory experience and emotional state[9] –the conception of mind we associate with 'higher' organic life rather than mere mechanical devices; no one has yet demonstrated that computers 'experience' fear, anger or indeed any other emotional state. Attending to stimuli need not entail awareness of that stimuli, however. While awareness always co-occurs with attention, attention may not imply concurrent awareness (Schmidt, 1995, 2001). The child glancing at a book cover might notice (i.e. display awareness of) the general design of the layout and form, yet remain unaware of the author's name despite that this also falls within his or her visual field; or perhaps s/he attends to, and exhibits awareness of a particular cover detail –the barcode, say– while

[9] The terms awareness and consciousness are synonyms.

remaining unaware of the general composition.

In studies dating from the mid 1990s, Schmidt argued that the attention/awareness distinction implied two critically dissociable second language appropriation mechanisms: (1.) Conscious (arising under the +attention +awareness state) and, (2.) subconscious (arising from a state of +attention –awareness) i.e. from attention alone. Schmidt (2001) defined awareness as consciousness in the limited 'focal' sense of whatever amounts to the *object* of one's attention at a particular time point, as opposed to the *global* sense (see e.g. Truscott, 1999) of consciousness of the stimulus generally i.e. holistically[10] (Returning to the book cover illustration, above, the global sense corresponds to awareness of the cover as a whole, and focal, to a detail). For short duration instances of focal awareness Schmidt (1995) coined the term *noticing*, arguing that a learner *notices* a target language detail whenever s/he consciously registers that detail during a moment of language exposure. *Noticing* behaviors in the classroom might include, for example, a student's realization that a word carries an alternative meaning given a change in the speaker's pitch, that verbs undergo inflection for tense, or that nouns carry markers for plurality (see Schmidt, 1995, 2010).[11]

Schmidt (1995, 2010) claimed two contributions of *noticing* to L2 development.[12] First, the process alerts students to language features they have yet to master; in this role it serves as a 'pointer' to language structures worthy of further (formal) study. Second, and critically, *noticing* functions as the preconditional behavior –"the necessary and sufficient condition," Schmidt (1995) argues –from which all learning necessarily proceeds. Truscott (1998) termed this the 'strong' form of the *noticing* hypothesis in contrast to the 'weak' form that proposes *noticing* serves a purely 'learning facilitative' function. To distinguish between the *products* of conscious and subconscious language appropriation, as opposed to the *processes* from which each derives, Schmidt (1995) introduced the terms *learned* (i.e. explicit) knowledge, and *acquired* (i.e. implicit) knowledge. Citing brain lesion studies, he draws

[10] Global awareness is presumed preconditional for both conscious and unconscious language appropriation.
[11] These examples are all of "focal awareness," awareness of a particular aspect of the stimuli. It is this, rather than global (general) awareness, which equates with *noticing* (Truscott, 1998).
[12] Schmidt's (1995) noticing test: If a child can *recall* a language feature at the time of learning then s/he has noticed it.

attention to qualitative dissimilarities between the two knowledge types noting that each uniquely occupies one or the other of the two human memory stores –conscious (i.e. learned) L1 or L2 understandings residing in the declarative store, sharing this with non-language learned facts and details, and acquired knowledge, the procedural store.

As a repository of 'facts' and 'rules,' learned language exhibits the properties of declarative knowledge generally: this includes, for example, the potential for expression in speech or writing (we can explain to others what we have mastered), gain from formal or self-study (teachers can transmit 'facts' to students), and availability for introspection –e.g. a child/adult might self-analyze what s/he consciously understands. Paradis (2009) describes a defining feature of declarative knowledge as 'arbitrariness.'[13] That the string [c-a-t] refers to a four-legged mammal, for example, and not a precious metal or article of cutlery represents an arbitrary association. Nothing of [cat] hints at what the term denotes. Other examples might include telephone numbers, the name of a lottery winner, or the favorite color of a randomly selected child.

For EAL learners, the subject of the current study, learnt rules (i.e. declarative memories) serve two main functions. First, they give rise to a subset of language competences that develop only from learning behavior irrespective of first language of L2 learner status. Second, they allow for generating well-formed written or spoken output should the EAL learner's implicit (acquired) competence currently prove insufficient for the task (see Table 2.1, p.23). Depending upon chronological age and cognitive maturity, learnt rules will derive from either self-teaching, formal instruction or simply the act of *noticing*. A reader, for example, may *notice* that verbs carry an '*ed*' marker to denote 'completion,' before going on to formulate a simple rule to capture the new insight –perhaps a note of some sort to append 'ed' to verbs to indicate a finished action. Conscientious EAL students may often gain extensive rule knowledge from just such reflection and language study.

Aspects of language *necessarily* learned –i.e. those non-acquirable, whether native speaker or

[13] An example of a rule governed association would be the rules governing aspiration of /p/, the circumstances in which this occurs being readily predictable from context, as opposed to purely arbitrary.

otherwise– include the correct association of a word (i.e. its phonological or orthographic form) with its meaning (Lum & Kidd, 2012), the appropriate number of arguments accompanying a verb (e.g., that *hit* takes a direct object), the correct occurrences of irregular past-tense forms (teach → taught), and how to interpret idioms with meanings non-deducible from the sum of their parts (Ullman, 2001, p.233). More comprehensive listings appear in Hartshorne and Ullman (2006) and in Paradis (2009). To Schmidt (1995), learned language competencies will emerge without formal study as such, the act of *noticing* sufficing in itself. Deliberate study of the sort students might undertake in schools or language classes, for example, served the 'academic' purpose of taking language learners beyond declarative knowledge gains and into a state of deeper technical L2 familiarity that Schmidt (1995) labels *understanding.* The term denotes the type of linguistic familiarity that teachers and academics draw upon to explain intricate language peculiarities and which Krashen (1988) has long referred to as monitor knowledge.

Schmidt (2000), like Paradis (2009), describes acquired knowledge as subconscious rule-like, computational procedures that derive from a subconscious aggregation process[14] that operates by summing features of language intake[15] to which it exhibits sensitivity; for a detailed discussion, see Ellis (2002). Language users experience this knowledge as hunches or intuitions: a sense that a clause appears ill-formed, for example, despite ignorance of a 'grammar book' rule violation, or a 'feeling' that certain letter strings seem impermissible (e.g., 'xc,' 'zx,'), or that a word collocates with another despite that one cannot offer a rule based explanation (e.g. that *torrential* 'goes with' *rain*, but not with *snow*). For native English speakers[16] implicit knowledge generates correct syntactic and phonological forms from simpler base elements ('loved' from 'love;' 'greater' from 'great'), ensures syllables receive the appropriate pitch and stress (Ellis, 2000; Paradis, 2009, p.54), that regular verbs

---

[14] This process involves tallying certain details (often of a language detail not consciously accessible to the learner) from the language input and, from these, building sub-conscious rule like understandings (see Paradis, 2009; Ellis, 2005).
[15] Intake refers to what the language user subconsciously abstracts from the input (speech and writing) s/he encounters.
[16] For ESL students, as we shall see, acquisition difficulties may oblige the child to employ conscious rule proxies for implicit knowledge.

(but not irregular –these a child must learn) receive the proper inflection for tense, that appropriate articles appear before nouns, that word order choices exhibit contextual appropriateness (Pinker, 1999), and that speakers construct permissible sound combinations (Ullman, 2004, p.245). Just as declarative knowledge has its own defining properties, so too, acquired. Acquired knowledge always remains inaccessible to introspection. It is inexpressible through writing or speech and does not arise from deliberate learning efforts.[17]

Despite qualitative differences, the two knowledge types –acquired and learned– operate cooperatively to provide their own unique contributions to language 'output' well-formedness (Ullman, 2001). For native English speakers and EAL students alike, learned knowledge generates 'output' features that arise from applying consciously gained capacities/rules (p.20), while acquired supplies those language features that users generate subconsciously (Ullman, 2001). For EAL students, learned rules also serve an important subsidiary function: they act as substitutes for any implicit understandings that as yet remain unacquired. Adult EAL learners will often rely on just such rules to compensate for the implicit learning difficulties associated with advancing chronological age (see Section 2.5.2). The rules perform a 'fill in' function until such time, if ever, the corresponding implicit understanding becomes secure.

Paradis (2009) describes the declarative and procedural memory stores as anatomically distinct, each susceptible to different types of injury, impairment and performance limitations. The declarative store 'resides' in the temporal cortex (Ullman et al., 1997), while implicit knowledge representation lies in the cerebellum, putamen, caudate nucleus, and the motor cortex. The manner of gain determines the memory store that a language competence occupies. Competencies gained consciously (e.g. under a state of '+ attention + awareness') subsist in the declarative store where they define what Paradis (2009, p.17) calls the 'vocabulary.' Aspects of language competence gained subconsciously i.e. implicitly (via the ' + attention –awareness' condition), on the other hand, occupy the procedural

[17] One can describe the output of implicit knowledge (i.e., what it supplies). We cannot, however, express explicitly what we do not have conscious awareness of.

memory store and comprise the 'lexicon.' A list of the main distinctions between the two terms appears in Table 2.1 below.

Descriptive label for knowledge type	Terms for manner of appropriation	Memory store
lexical	1. Subconscious (+ attention-awareness) 2. Implicit	Procedural
vocabulary	1. Conscious (+attention +awareness) 2. Explicit	Declarative

Table 2.1: Types of language knowledge by property and manner of appropriation.

Although subsisting as dissociable and autonomous structures, the learning and acquisition systems impact upon one another subtly and indirectly (Ellis, 2005). A student might, for example, draw upon acquired knowledge to self-generate clauses that s/he then consciously analyzes to derive declarative language understandings (i.e. rules) that subsequently become part of his or her declarative store. Learning, conversely, impacts upon implicit system operation whenever the learner '*notices*' L2 input features that the implicit system currently lacks the capacity to process. Repeated acts of such noticing sensitize the implicit system to 'new' language features through a *retuning process* that potentially allows for L2 competence gains in the manner they arise during L1 development (Ellis, 2005; see also Section 2.5.2). Despite such interactions, however, the output of the two systems nevertheless remains qualitatively different (see Krashen, 2004; Rodrigo, 2006). EAL students will fail to gain implicit competencies from practicing pedagogical rules to the point of rapid execution since those competencies never amount to the "speeded up" application of declarative understandings

(Ellis, 2005, p.333; Paradis, 2009, p.86). Likewise, implicit knowledge will not evolve into declarative, the latter arising from –and *only* arising from– a conscious learning effort. This does not mean that learned knowledge fails to generate speech and writing of comparable ‘naturalness’ and grammaticality to that of language ‘acquirers’ (Paradis, 2009) –at least under circumstances conducive to conscious rule application. Diligent adult EAL students may ‘come across’ as authentically native like even to English mother tongue speakers. The mental processes they evoke to supply that same output nevertheless always remain fundamentally dissimilar -on the one hand, declarative understandings (*learners*), and on the other (*acquirers*), subconscious, automatized, implicit skills.

Given two children, one an EAL student and the other an English native speaker, learned pedagogical rules rarely make comparable contributions when generating the same surface structures. For a primary aged Year 4 English monolingual child having acquired English in the home environment, the default source of language competencies appears in list form in Table 2.2. The following table, Table 2.3 lists the same language capabilities, manner of gain, and memory storage locations for a hypothetical Thai L1 EAL student attending the same Year 4 class but for whom regular exposure to English began at age 8, a few months previously. Whether this second child need learn rule substitutes for competencies native speakers acquire implicitly depends upon how effectively his or her acquisition system functions given the lesser sensitivity to non L1 language features accompanying optimization for L1 input processing during early childhood (Bley-Vroman, 2009). Should optimization have rendered the system unreceptive, or minimally receptive, to L2 intake then acquired competencies either fail to emerge altogether (Ellis, 2002) or will do so relatively slowly and erratically. For the hypothetical Thai L1 child’s implicit ‘learning’ to proceed, one or both of two conditions need apply: either that system must retain some residual receptivity to non-L1 features or, and following successful sensitization, the system gains sensitivity to language features it had become unreceptive to (Ellis, 1994, 2008). Ellis (1995) cautions that sensitization (i.e. ‘*retuning*’) does not arise without *noticing*, however, and may require a sustained conscious learning effort (see Section 2.4.2).

The common feature in Tables 2.2 and 2.3 lies in the designation of ‘word meaning gains’ as the product of conscious learning (+attention + noticing) behavior. Whether learning English as an additional language or mother tongue, building associations between phonological (or orthographic) forms and meanings minimally requires *noticing* on the learner’s part (see e.g. Ullman, 2001).

Manner of learning	*Descriptive term*	*Language gains*	*Where stored*
1. Conscious	*Learning*	Word meanings, meanings of idiomatic expression	Declarative memory
2. Non-conscious	*Acquisition*	Grammatical properties of words, including: 1. Number of objects a verb takes 2. Skills underlying correct Articulation 3 Skills underlying correct intonation 4. Skills underlying correct word stress 5. Collocation knowledge 6. Syntax 7. Morphology	Implicit memory

Table 2.2: Types of learning and their products (L1).

Manner of learning	*Descriptive term*	*Language gains*	*Where stored*
1. Conscious	*Learning*	Word meanings, idiomatic expressions, and the grammatical properties of words	Declarative memory
2. Conscious (i.e. learning) or non-conscious (acquisition), depending upon whether the implicit system is receptive to acquiring relevant L2 language features.		Grammatical properties of words, including: 1. Number of objects a verb takes 2. Skills underlying correct articulation 3. Skills underlying correct intonation 4. Skills underlying correct word stress 5. Collocation knowledge 6. Syntax 7. Morphology	Implicit or declarative memory

Table 2.3: Types of learning and their products (L2).

2.3 Explicit, implicit, and incidental vocabulary acquisition

Building upon the discussion of *noticing*, the current section moves on to define the types of vocabulary learning processes that occur during reading experiences; namely: (1.) explicit, (2.) incidental, and (3.) implicit (Rieder, 2003). The section limits itself to examining these terms within the contexts in which they continue to attract attention in linguistic research: (1.) As mental states (i.e., the affective disposition that an encounter with a novel word induces), and (2.) as learning behaviors (i.e., how a child responds to a new word at the moment such encounters arise).

Explicit vocabulary learning (as a mental state + behavior)

As a mental state, explicit learning refers to an intent or motivation on the learner's part to assign meaning to an unfamiliar or partially known word for the purpose of retaining that understanding for future receptive and/or productive use (Ellis, 1995; Nation, 2001).[18] The behaviors this state commonly evokes during recreational reading fall into two categories:[19]

1. Meaning-derivation: This includes, for example, consulting thesauri or dictionaries, guessing from context (GFC), seeking clarification from teachers, or attending to marginal glosses (see Schmitt & Schmitt, 1993).

2, Consolidation strategies: These aim preserve a permanent record of the new meaning in the long-term memory store. Such strategies include semantic mapping (Heimlich & Pittelman, 1986), verbal repetition (Shiwu, 2005), the keyword technique (Atkinson & Raugh, 1975), and feature grids (Anders & Bos, 1986).

Examples of explicit learning in practice:

a. A young girl reading a book for pleasure comes across the word 'elevator.' She pauses, thinking to herself, *this would be a useful word to know. I'll write it down!* (consolidation strategy).

b. A young boy comes across the word 'canopic.' He thinks, *I've heard that word before.* The child then attempts to infer meaning from context (meaning derivation strategy).

Incidental learning (as a mental state + behavior)

The mental state operative during incidental learning consists of the absence of any intention to learn whatever the child nevertheless successfully learned from a particular behavior (Rieder, 2003).

---

[18] In Hulstijn's (2001) terms, explicit learning amounts to "any activity geared at committing lexical information to memory" (p. 271).
[19] Strategy: "Any activity the learner consciously chooses for the purpose of regulating their own learning" (Griffiths, 2008, p.87).

Either the child remained intent on learning something other than what s/he indeed gained or, albeit uninterested in learning as such, yet unintentionally retained some language understanding from the reading experience. Incidental learning itself falls into two sub-types: (1.) explicit and (2.) non-explicit.

Explicit incidental learning denotes a state of intention to assign meaning to an unfamiliar word along with a willingness to apply strategies to achieve that end.[20] The reader has no interest in word gains as such, nor indeed any desire to retain a derived meaning for longer than suffices to interpret the text currently engaging his/her attention (herein lies the contrast with explicit learning which involves just such a retention goal). The typical classroom behaviors incidental learning induces include consulting dictionaries and thesauri, requests to teachers for word meaning clarifications, and efforts to derive word meanings from contextual clues. Any meanings that persist in long-term memory amount to the unintended outcomes of the reading opportunity (Rieder, 2003).

Non-explicit incidental learning as a mental state denotes awareness of a novel word meaning from a word encounter which, for the particular reader and given his or her background knowledge and language skills, suffices to ensure the meaning becomes readily apparent. The novel word itself evokes no desire on the reader's part to establish the meaning as such (the context provides such clarification); nor does having experienced the word induce an effort to preserve the presumed meaning for future productive or receptive language use.[21] The behavior most associated with non-explicit incidental learning consists of pleasure reading without accompanying forms of teacher-fronted, post-reading, accountability (e.g., tests, or instructor requests for oral summaries etc.). Any unfamiliar words that indeed become additions to the vocabulary represent the fortuitous outcome of pursuing a pleasurable reading task (Hulstijn, 1989).

[20] "The special thoughts or behaviors which individuals use to comprehend, learn or retain new information" (O'Malley & Chamot, 1990, p.1).
[21] Were it to induce either, or both, then learning would fall under the explicit label.

The two forms of incidental learning each arise from different levels of awareness (i.e. consciousness). For explicit incidental learning, awareness lies along a continuum bounded by marginally more than *noticing*[22] through to the deep-level awareness ("understanding" in Schmidt's (2001) terminology) that strategy selection and application necessarily entail. For non-explicit incidental learning, conversely, awareness amounts merely to *noticing* of a form-meaning pairing. The learner does not, that is, attend to the pairing other than to incorporate the presumed meaning into his or her understanding of the text. In terms of outcome, explicit incidental learning proves the more productive of the incidental learning forms for amassing word meaning gains per unit of reading time. It contributes rather less to long-term vocabulary uptake, however, since it typically represents the less frequent behavior during pleasure-reading sessions (Nagy et al., 1985).

Examples of incidental learning in practice

a. The child comes across the colloquialism *raining cats and dogs* and reflects upon what this means. S/he asks a friend sitting at a table nearby who explains that it refers to heavy rain (explicit incidental learning).

b. The child reads the word *mudblood* in a novel about a child wizard. The author thoughtfully provides a definition of the term in the following sentence and then again at various other points in the text. The reader retains the meaning of that word despite having no interest in doing so –it so happens s/he dislikes novels concerning young wizards! (non-explicit incidental learning).

<u>*Implicit learning (as a mental state + behavior)*</u>

Acquisition (or implicit learning) arises during a state of unawareness (zero *noticing*) of the features of the input from which a competence arises (Carroll, 2006; Schmidt, 2010, p.9)[23] albeit the

[22] If the learner did no more than *notice* a form–meaning association, then we have non-explicit incidental learning.
[23] As Caroll (2006) explains, most of what is input consists of mental constructs that exist in the mind and not overtly in language output as such.

'learner' exhibits global awareness of the stimulus as a whole (Truscott, 1998). Underlying the process is the language user's subconscious sensitivity to co-occurring features of language intake, together with what Ellis (2002) calls a 'dumb' instance based aggregation mechanism that tallies certain regularities of those features to establish associations between them and the constructions in which they arise. The result of this tallying consists of subconsciously 'familiar' memory chunks of "contiguous components" (Ellis, 2005, p.307) which, in turn, serve as elements of 'rule-like' implicit understandings. Tondaki (2015) cites such tallying as explaining our sense of collocation. While we readily acknowledge the phrase 'pretty *penny*' as acceptable, we just as readily reject 'a pretty *pound'* as somehow odd or anomalous. Ellis (2002) describes this subconscious familiarity with the frequency of feature co-occurrences as the essence of implicit human language competence. Outside of language proficiency, implicit knowledge underlies performance of various skilled tasks: playing tennis, riding a bicycle, typing, and so on (Ullman, 2001).

Examples of implicit learning in practice:

a. A learner might attend to 'the gist of a conversation,' gaining something of collocation, prosody, or word class while having no intention or aim to do so (Hasher & Chromiak, 1977, p.173).

b. While reading, a child gains an understanding that the definite article (*the*) precedes unique objects, as in *the* moon, *the* earth.

The several manners of learning/acquisition (above) may co-occur during the same reading session. A pleasure reader might pause occasionally to learn an unfamiliar word just as those who tend to ponder over each new vocabulary term may do so rather less should the text prove sufficiently engrossing. Since the several learning/acquisition types make quantitative and qualitatively different contributions to vocabulary development, a child's rate of word-meaning depends upon the time allocated to each learning type. Explicit learning proves most 'productive' for the time expended, followed by explicit incidental, and then non-explicit incidental. Implicit learning functions as an ongoing 'background' process to the other learning forms, contributing minimally to the stock of word-meaning associations, such associations arising almost exclusively (Williams, 2005) from *noticing* matches between orthographic forms and associated percepts (Ellis, 2005; Paradis, 2009; Ullman, 2004).

2.4 Spaced learning

Academic interest in spaced learning (Section 1.2) has historically centered upon quantifying economies in instructional time from controlling the duration and frequency of inter study intervals (ISIs) between learning opportunities. Early research identified a spaced (or distributed) learning advantage that expresses itself in more learning per unit of time from study periods interspersed with ISIs than from a single uninterrupted session (Dempster, 1988, p.627; Ebbinghaus, 1895). The effect has proved both replicable and robust (Dempster, 1988). It arises under intentional as well as incidental learning conditions (Challis, 1993), from encounters with identical or non-identical stimuli (Vlatch, Sandhoffer & Kornell, 2008; Vlach & Sandhoffer, 2012), and among children as well as adults. The time savings from spaced learning, moreover, emerge from a broad range of task types. Keller, and Atkinson (1967), for example, reported a spaced learning advantage from an experimental spelling program; Donovan and Radosevich (1999) cited time savings in motor skill performance while Dempster (1988) and Krashen (2004) described impressive learning outcomes from a 'spaced' vocabulary program. Evidence for a spaced learning advantage in paragraph recall tasks appears in Noel and Sawyer (2003), while Kornell and Bjork (2008) found a distributed learning advantage under inductive learning conditions, as did Kornell, Castel, Eich, and Bjork (2010) in a replication of that study. Other research has reported time savings from skill learning activities such as, for example, mirror tracing and video game mastery (Donovan & Radosevich, 1999).

Despite a bias towards laboratory investigations (section 1.2), several studies report school-based findings. Reynolds and Glaser (1964) claimed time savings from a secondary level biology course incorporating distributed learning tasks. Rea and Modigliani (1987) noted impressive results among primary school children from a spaced spelling and multiplication program while Bloom and Shuell (1981) recorded a substantial 35% increase in word gains under distributed learning conditions among high school students studying French as an L2.[24] Knowledge retention, too, may prove rather more durable given spaced learning opportunities. Sobel, Capeda, and Kapler (2011), reported superior vocabulary retention among 39 Grade 5 pupils from spaced instruction (this involved presenting slides, oral practice along with paper and pencil assessments) from tests conducted one week after the last tuition session. The authors conclude that optimal intervals between study periods facilitates vocabulary uptake "in applied settings" with "middle-school-aged children" (p.763).

To realize spaced learning gains, ISI duration need not lie within narrowly circumscribed boundaries (Thalheimer, 2006, p.16). In a school-based study involving Year 1 children, Seabrook, Brown, and Solity (2004) compared learning outcomes from 'long' study sessions (of phonics) with those from 'shorter,' the manner of teaching held constant under each instructional condition. Assessments undertaken upon course completion revealed a six-fold increase in gains of grapheme/phoneme associations among children exposed to shorter but more frequent learning opportunities. Though conceding a need remained for further research –the study involved just 34 students– the authors point to the study's strong ecological validity: The investigation took place in an authentic school setting among pupils performing to age-related expectations and with regular classroom teachers providing instruction. The study questions the efficacy of 'standard-duration' lessons (40 minutes or so) suggesting shorter, more numerous, learning sessions noting that this may have minimal, or zero, cost implications.

In a meta-analysis of previous research, Janiszewski et al. (2003) identified four such factors that moderate the effect size from spaced learning opportunities:

---

[24] This from testing which took place four days after training ceased.

- In general, the longer the ISI the greater the effect size, the ideal ISI "being the longest ISI before items are forgotten" (Bahrick & Phelps 1987, p.370).[25] Since most forgetting invariably arises soon after the initial learning (Pimsleur, 1967), this implies short intervals initially, giving way to longer as the product of learning becomes more secure in memory (Schacter, 2001). Pimsleur's (1967) learning schedule, Mondria and Mondria-de Vries' (1994) 'hand calculator,' and Wozniak's *SuperMemo* (Biedalak & Wozniak, 1996) incorporate just such ever-increasing time spans between learning occasions.

- Stimulus complexity (semantically complex stimuli evoke a larger effect size than structurally complex or simple stimuli).

- Learning type (intentional learning ensures a larger effect size than incidental).

- Complexity of intervening material (intervening material that is semantically complex is associated with larger effect sizes than intervening material that is structurally simple).

Despite a wealth of compelling evidence, and urgings from researchers (e.g., Seabrook et al., 2004; Thalheimer, 2013), few schools in the U.S. or U.K. currently exploit the spacing effect to optimize classroom instruction. Reporting of American teachers, Dempster (1988) described attitudes to spaced learning as ambivalent, noting that many classroom instructors claimed the benefits remained unproven or doubting if tangible gains arise in 'applied' as opposed to clinical settings. Dempster (1988) attributed these misconceptions firmly to researchers' failure to build upon findings from earlier studies. Without chains of linked research, he argued, the relevant literature struck educators as both "ahistorical" and unpersuasive. Carpenter, Cepeda, Rohrer, Kang, and Pashler (2012, p.375) noted

[25] "As an empirical rule, the generalization seems to be that a repetition will help most if the material has been in storage long enough to be just on the verge of being forgotten" (Crowder,1989, p.49).

similar skepticism among educational stakeholders more generally, including course designers and material writers. Citing Rohrer and Taylor's (2006) observation that few commercially available mathematics programs acknowledge time savings from spaced topic presentations (see also Pashler et al., 2007) Carpenter et al (2012) concluded that education providers remained ignorant of distributed learning advantages (p.375). Students, too, would seem no less unaware. Drawing upon records of graduate participants' study habits, Cohen, Halamish and Bjork (2013) found little indication this population realized the potential time savings from spaced study schedules. The cost in terms of lower grades remains unclear. For middle-school language students, however, Lindsey, Shroyer, Pashler and Mozer (2014) claimed that researcher-designed, personalized, distributed learning review programs could result in a substantial 16.5% addition to topic knowledge retention over that derived from massed study and 10% more retention than observed from "one-size-fits all" (p.6) spaced review programs.

2.5 Moderating factors

To the technical determinants of spacing effect efficacy[26] that Janiszewski et al. (2003) identify in their meta-analysis, Dempster (1988) noted additional factors that moderate children's learning under spaced learning conditions more generally. These factors fall into two categories (1.) learner attributes, and (2.) classroom conditions. The first category includes the cognitive maturity children bring to the vocabulary learning task, the learner's chronological age, how developed the child's Theory of Mind (ToM), and his or her first language. In the second category lies the child's school environment and its general conduciveness for language learning/acquisition experiences. This category includes most notably the quality of inter-personal relationships between EAL learners, their teachers and student peers.

This current section explores each category of factor in turn. Findings from this discussion will assist in piecing together a comprehensive account of how young EAL children gain word meanings

[26] These include the number of ISIs, their duration, the total length of the learning session etc. (see Janiszewski et al. 2003).

from RR experiences (see Section 2.8).

2.5.1 Moderating Factor 1: Cognitive maturity

Children begin to speak recognizable words at around the age of 12 months (Bloom, 2000). By age three, a child knows over 1000 word families (Hart & Risley, 1995) and from ages 5–11 typically learns around eight words per day (Goulden, Nation, & Read, 1990), corresponding to around 2,900 per year. To make such gains children draw primarily upon two sources: (1.) an evolving understanding of syntax, and (2.) an emerging Theory of Mind (ToM). From syntactic knowledge children identify part of speech (whether a word falls into the category of noun, verb, or adjective etc.), likely animacy/inanimacy, and the contexts in which a word occurs, among others. Hearing "*The feb chased the cat*" the child deduces that *feb* stands as a noun (the definite article proves helpful here), that it exhibits 'life' and displays qualities of aggression or mischievousness.

The term Theory of Mind refers to the human sensitivity to understand and interpret the emotional states, beliefs, intents, and desires of others. This sensitivity allows children to associate novel words with the speaker's referential intentions by reliably interpreting facial expressions, vocal stress, pitch, and human behaviors more generally. From a stern "*No!*" the child deduces prohibition; an alarmed "*Stop!,"* urgency, and from "*Excellent!*" that a behavior meets with peer approval. As ToM and syntactic knowledge develop into the adult form (see Wimmer & Perner, 1983), so metacognitive skills increasingly predict how successfully word meaning derivation and consolidation in memory proceed. Since these skills derive from life experience and education (Bloom, 2000), word learning capacity typically improves with advancing chronological age. The manner in which 'young' and 'old' gain word meanings, however, remains qualitatively the same. A graduate learns vocabulary more ably than a nine-year-old not from accessing a learning mechanism unavailable to the child but from applying more extensive metacognitive skills to derive form/meaning associations and establish these in long-term memory.

Bloom (2000) identified two difficulties with alternative, associative, i.e. non-ToM-based models. First, children only rarely construct the false form-meaning pairings we might reasonably expect were such learning the source of gains; for example, a child does not typically confuse *dog* with *foot* (though both in a sense occupy the floor), *plate* with *spoon* (though both rest on tables), or *shoe* with *sock*. Second, the aggregation process underlying associative learning (p. 30) leaves unexplained children's accelerating vocabulary uptake with ongoing maturation. That teenagers learn more ably than ten-year-olds, and ten-year-olds more so than five-year-olds suggested a learning mechanism that builds upon previous knowledge as opposed to mechanistic tallying from multiple stimuli exposures (Smith & Yu, 2008, p.34). To Bloom (2000), ToM also explained aspects of children's vocabulary understanding for which associative learning had long failed to provide plausible answers (Bloom, 2000). With deduction, as opposed to 'tallying,' the source of word meaning gains, the *dog/foot* type mislabeling, for example, now predictably becomes relatively infrequent while documented cases of word meaning gain from just one or two informative exposures (Dollaghan, 1985), though rare, notably less surprising. The deductive based word learning hypothesis also offered a plausible explanation of the poor vocabulary uptake following impairment of neurological structures that sustain learning as opposed to acquisition. The degraded capacity to *notice* associated with anterograde amnesia always ensures difficulty matching phonological forms with meanings despite that patients acquire implicit competencies as ably as those unafflicted (Paradis, 2009). Among autistic children, too, for whom the ToM remains undeveloped, successful learning of form/meaning associations arises only after much effort as does 'uptake' of declarative facts/details in general (Baron-Cohen, Leslie, & Frith, 1985; Bloom, 2000).

Bloom (2000) described vocabulary learning as "continuous" meaning that the rate at which learning progresses rarely deviates over time. He rejected McCarthy's (1954) claim of a childhood vocabulary growth spurt (beginning around age 2) on methodological grounds pointing to a misplaced focus upon rate of growth[27] as opposed to change in rate, proposing the latter as the more relevant

[27] The assumption Bloom (2000) objects to is that evidence of a "spurt" comes from a child simply attaining a certain rate of growth as such (words per day, week, month, etc.).

measure. Other investigations to founder on this same oversight include Dore (1978), McShane (1979), and Behrend (1990), all of which Bloom (2000) identified as ‘growth spurt’ supportive. From reexamining the available evidence, Bloom (2004) concluded vocabulary development typically reveals itself as both continuous and incremental, citing Ganger and Brent’s (2004) failure to detect inflection points (i.e. sudden deviations from trend suggestive of a spurt) in the growth curves of 38 one and two-year-olds. This, the ‘incremental hypothesis,’ receives further support from Moore and Louis ten Bosche (2009) who noted an absence of abrupt departures from vocabulary developmental trends in the logistic growth curves of their 1,800 child participants, aged between 8 and 30 months.[28] Despite that some children indeed displayed short duration learning ‘spikes,’ only a few demonstrated an abrupt onset and sustained departure from previous growth that the spurt hypothesis predicted. The study concluded, as had Bloom (2004), that evidence for a spurt seems “slim,” arguing that a Gompertz function offered a “satisfactory, ecologically-motivated, model of lexical growth from birth to young adulthood”[29] (p.5).

A still contentious question concerns whether ToM-based models adequately explain just how a child correctly identifies a word meaning from among the several possibilities a context might imply. Quine (1960) termed this the *referential indeterminacy* issue, asking how children could deduce from the adult’s “Look! We have a *gavagai* in the garden!” a reference to the rabbit as a whole, the carrot lying nearby, or the rabbit’s fluffy tail. The answer, according to Markman (1994), lies in innate understandings that predispose children towards accepting certain word/meaning pairings as more plausible than others. These understandings include, for example, the *whole-object assumption*, a bias that prompts a child to assign novel words to objects as opposed to their constituent parts. The ‘toddler’ hearing *“house*” as his mother points to a photograph of the family home therefore more readily associates the term with the building as opposed to the windows or doors. A second bias (Markman,

[28] A logistic growth curve is an S-shaped (sigmoidal) curve that models functions that increase gradually at first, then more rapidly, slowing down to leveling off at a maximum value.
[29] A Gompertz function: a mathematical model for time series data where growth is slowest at the start and the end of a time period.

1994), the *taxonomic assumption*, inclines learners to assign meanings based upon taxonomic rather than thematic relations. A child familiar with the term *rabbit* (i.e. the word exists in his/her oral vocabulary) therefore more likely associates this, the known referent, with similar animals (furry, long-eared i.e., of the same taxon) rather than with *hutches, carrots, water trays* or other thematically related objects. Markman's (1994) third bias, *mutual exclusivity* (see Clark, 1983) refers to children's tendency to assume an object has its own designative term unique to that object; an object cannot be both a table and a chair, or cup and a saucer, but, rather, one or the other. The bias would explain why learners typically assign unique words to unknown entities rather than 'extend' a familiar word's referential scope to include that entity. A child observing the teacher holding up a *book*, a known word, while uttering '*red'* (unknown) therefore concludes that *red* denotes a quality of the book rather than standing as a synonym for the familiar term. De Witt (1994) notes that mutual exclusivity does not frustrate L2 vocabulary gain, however, since the bias operates *within* but not *across* languages. Monolingual and bilingual children readily accept alternative labels for the same object so long as they can assign one to the L1 and the other to the L2 (Au & Glusman,1990).

Markman (1994) believed the biases expressed general principles applicable to non-language domains including perception and numerical cognition (p.241). He cites Shipley and Shepperson's (1990) finding of pre-school children's tendency to sum partial objects as if complete entities: a fork broken into two pieces became two whole forks in an apparent application of the *whole object* assumption. Other evidence includes Flavell's (1988) work with 3-year-olds from which derived the 'one identity presumption,' a more expansive predisposition than mutual exclusivity that claims to explain aspects of visual and conceptual perspective processing. More recently, Sloutsky and Fisher (2012) have drawn parallels between the *taxonomic assumption* and inductive projection –i.e. a sense that the same categories or classes of objects share common attributes (see e.g.). That object 'x' has claws and fur, for example, would inductively suggest an association with object 'y' that possesses the same attributes, as opposed to 'c' that displays neither. Markman's (1994) essential premise that heritable understandings moderate human stimuli processing still goes largely unchallenged. In the late 1950s Chomsky (1957) demonstrated that innate restraints limited the class of permissible human grammars providing a long sought answer to how children acquired language despite grammatical errors in caregivers' speech. In the animal learning literature, too, innate predispositions have offered convincing explanations of behaviors ranging from preparedness to respond to certain stimuli (e.g., Shettleworth, 1972) and migration patterns in birds to defensive postures in wild dogs.

Whether ToM or associative learning better explains word gains, vocabulary uptake during RR always occurs within the context of the broader cognitive challenges reading entails. The conventional view (see e.g., Adams, 1990) sees children assigning meanings to words in one of two ways: either via a conscious, serial, left-to-right decoding/blending effort (the so-called non-lexical, or sub-lexical reading route) or, subconsciously (the lexical route), such that the visual image of the word evokes a pre-stored meaning representation in memory (Snowling & Hulme, 2008). The non-lexical route proves the slower, more effortful, and more cognitively demanding of the two (Adams, 1990). It is the process emergent readers employ and, on occasion, skilled readers should they encounter a novel word of sufficiently unfamiliar orthography[30] that lexical reading remains unavailable. To read non-lexically the child methodically works from left to right through the string of interest, recalling and applying pre-learned grapheme/phoneme correspondence rules to match each letter in turn, or letter combination, to associated phonemic expressions ("t" becomes /t/; "ee" becomes /i:/, and so on). The reader now stores each recovered sound element (or phoneme) in short-term memory before moving on to analyze the next letter(s) of interest. Once s/he has processed the string in its entirely in this manner, a blending operation follows during which the reader amalgamates the recovered phonemes into a single phonological representation (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). S/he completes the word identification effort by attempting to match this representation with a word in his or her oral lexicon (Adams, 1990). Should no match exist, a child will typically 'skip' the word or go on to employ word meaning derivation strategies (Adams, 1990).

[30] "Familiar" means that the reader has encountered the word sufficiently that its phonological form will subconsciously evoke a meaning in memory.

The non-lexical reader's dysfluency, coupled with relatively slow text processing, stems from the attentional system's limited capacity along with the serial manner in which decoding and blending proceed (Adams, 1990). 'Limited capacity' (Broadbent, 1958) refers to the finite stock of attentional resources available for cognitively demanding tasks that acts as a bottleneck restricting the conscious effort a child can apply towards task completion. The non-lexical reader, having to execute multiple tasks (e.g. decoding, blending and textual interpretation), must optimally apportion attention to each (LaBerge & Samuels, 1974). Should a child over allocate to one, then s/he deprives attention from another, 'unbalancing' the reading effort. S/he might, for example, decode well, yet comprehend poorly, or comprehend impressively but blend slowly and with difficulty. The serial manner in which decoding proceeds compounds the difficulties, requiring from the child repetitive efforts to select and then apply grapheme/phoneme rules to each constituent letter of the string of interest (Huitt, 2003). Presented with the three-letter sequence d-o-g, the reader attends first to the 'd,' then to the 'o,' and finally the 'g,' in that invariant order. The child's reading rate in words per minute, along with the pleasure the text evokes, depends much upon how rapidly selection and application of relevant grapheme/phoneme correspondence rules proceed. Letters or strings that prove difficult to decode (perhaps the child's mastery of such correspondences remain insecure), will potentially obstruct recovery of the phonological form as a whole.

As Gough and Tumner (1986) explain, reading comprehension skill proves largely predictable from two factors: (a.) listening comprehension (LC), and (b.) decoding skill (D) i.e. (RC=LCxD). The more effort evoking the decoding, and the less the child understands of spoken language, the less s/he comprehends of a script. In this, *The Simple View of Reading*, Gough and Tumner (1986) argue comprehension develops in tandem with emerging LC or D. As readers transition from lexical to non-lexical reading (i.e. as D increases), so they free more attentional resources for textual interpretation and/or word-meaning derivation. Rather than attending to letters sequentially –working left-to-right through a word of interest– lexical reading allows for analyzing whole strings subconsciously and holistically employing a parallel processing behavior in which each letter assists in revealing the identity of its neighbors (Rasinski, 2004). This process depends upon an implicit understanding on the reader's part of weighted probabilities of letter co-occurrences and/or their corresponding phonemes: for example, that 'q' more likely precedes 'u' rather than 'l,' or 'h' follows 't' more often than it does 'z' (the learning mechanism consists of the tallying process noted in Section 2.2). Once the probabilities become secure, a reader who identifies the 'd' in [dog] will concurrently rule out a following 'x' while at the same time alerting himself/herself to the high probability of the following 'o.' An analogous process operating at the word level explains the likelihood of word occurrences: 'How do you <u>do</u>?;' 'A bird in the hand is worth two in the …' Such a 'holistic' reading behavior –the child processes words as single entities– means word length (i.e., the number of constituent 'characters') bears little relation to how rapidly a reader assigns meaning to phonological forms. A lexical reader recovers the meaning of an orthographically familiar *long* word (e.g., *responsible* –11 letters) almost as promptly as s/he does a familiar but *shorter* one (e.g., *hut* –three letters).[31]

The effort readers expend deducing word meanings and comprehending a script depends much on the author's choice of syntax and the complexity of the text's propositional content. Texts laden with non-intuitive concepts and complex sentence structures divert attention from word meaning derivation and towards resolving these additional barriers to understanding. Children typically, and inevitably, gain less vocabulary per reading time the more such diversion proves necessary. Overly simple texts, conversely, may merely consolidate the already familiar and fail to introduce novel language structures. To Krashen (2004), this leaves the ideal text for literacy gain as challenging but not unduly so –in other words, a text that draws a child marginally beyond the boundary of present competence but no further. Reading materials meeting this requirement satisfy what Krashen (2004) calls the i+1 condition of 'optimum difficulty' where 'i' designates current proficiency and '1' a small extension beyond this point. Among such texts, the most conducive for raising literacy competence and vocabulary knowledge consist of those that induce pleasurable reading experiences (Pilgreen,

[31] The reader cannot employ such reading for an unfamiliar word, however, because such a word has no established association between its orthographic form and meaning in long term memory.

2000). It falls to literacy teachers to ensure a ready supply of such optimal scripts (Krashen, 2004).

2.5.2 Moderating Factor 2: The first language

Many difficulties children experience learning a second language stem from the dissimilarities, or language distance (LD), between the child's L1 and L2. A Thai mother tongue speaker, for example, more readily gains other tonal languages than do students from non-tonal language backgrounds, just as English L1 speakers experience less difficulty mastering French, a language sharing Germanic and Latinate roots with English, than they do learning Mandarin which does not. The structural differences between languages fall under the category of either 'learned,' or 'acquired,' depending upon whether acquisition or learning represents the default route through which native speakers generate a particular structure of interest (Section 2.2). This gives rise to not one, but two, forms of LD –a 'learned' LD on the one hand and an 'acquired' on the other. The qualitative differences between learning and acquisition, as processes (see p.21), mean that a language distance of one type does not necessarily imply a comparable language distance of the other. A large declarative knowledge LD does not necessarily mean a similarly large implicit LD, nor a large implicit LD an equivalent declarative LD.

The learned LD remains 'small,' and learning comparatively easy, whenever the declarative knowledge sustaining one language also applies to the second. As the LD becomes larger, so the EAL student's learning challenge rises given the lesser opportunities to transfer what s/he knows of one language to his/her understanding of the second. A low learned LD typically equates with learning ease because 'carry over' opportunities obviate the need to relearn whatever the student has previously mastered during the course of L1 gain: i.e. the learner applies first language knowledge to the L2 (Ijaz, 1986). Should a child understand the concept of "volume," for example, from first language experiences s/he need only transfer that L1 meaning to the relevant L2 phonological (or orthographic) form to gain a second language 'vocabulary item.' Learning partial synonyms proves rather more cognitively challenging since the learner must consciously disambiguate the L2 meaning from its partial L1 equivalent and then commit to memory the relevant L2 meaning connotations. *Noticing*

L1/L2 usage discrepancies represents the minimum conscious effort for this process to succeed. Rather more productively and effectively, however, the learner might seek out any such distinctions perhaps through formal study. Sprouse (2006) terms this substitution behavior of L1 for L2 nuances of meaning, *relexification*.

For competencies children acquire implicitly (see Table 2.3 for examples), language distance denotes how common are those co-occurring features of L1 and L2 intake to which the implicit system responds (Ellis, 1994, 2005). The larger any such co-occurrence, the smaller the LD and the lesser the student's need to sensitize the implicit system to permit L2 competence gains. In 'low LD' cases the same tallying system (see p.21, for details) aggregates relevant common L1/L2 intake features to supply the statistical intuitions sustaining both L1 and L2 performance alike (see Section 2.2), explaining, in part, the previous example of the Thai speaker's relative ease learning Mandarin compared to his/her French L1 classmate. To the extent the implicit system proves unresponsive to L2 language features, given few instances of L1/L2 feature cooccurrence, the LD will prove larger, implying little or no likelihood of successful acquisition without implicit system retuning (see p.23). Appropriate responses to a high LD include (a.) consciously learning declarative rule substitutes for absent implicit understandings and/or (b.) engaging in behaviors (e.g. *noticing* and *noticing the gap*)[32] that bring the L2 feature to awareness and facilitating the retuning process from which acquisition may then proceed. Both options contribute to acquisition in like manner by raising non-salient L2 surface structure features to consciousness and establishing stimulus representations (SRs) to which the implicit system responds as it does to L1 intake (Ellis, 2002).[33] Noticing of the L1/L2 distinction from which the SR arose proves necessary only until such time that SR becomes securely encoded in long-term memory (Ellis, 2002, p.174). The rate at which acquisition proceeds thereafter depends upon how frequently the learner encounters exemplars of language, self-generated or otherwise, containing intake that retuning has now made accessible for tallying (Paradis, 2009, p.96). Given the aggregation process

[32] Noticing the gap refers specifically to conscious awareness of a difference (gap) between the two languages.
[33] See Section 2.3 above.

from which implicit knowledge gains arise (p. 21), the more numerous those encounters, the more rapid the competence gain. Any formal rules learners have previously mastered to compensate for implicit learning deficits remain available until acquisition of the L2 competence becomes secure, rendering themselves redundant for this purpose thereafter (Crowell, 2004; Ellis, 2002).

Bley-Vroman (2009) described the young child's implicit acquisition system as plastic, in the sense of 'malleable' or 'impressionable:' a system, in other words, fully receptive to *all* language intake and potentially allowing native competence in any language to which the child receives sufficient exposure. Once children begin to lose this plasticity they become progressively less subconsciously sensitive to surface structure features absent in L1 input. How successfully second language[34] mastery proceeds therefore depends much upon the implicit system's residual plasticity once regular L2 exposure begins together with how far optimization for L1 input processing has already advanced. The age beyond which L2 mastery requires conscious learning varies, however, depending upon the competence of interest (whether syntax, or pronunciation, for example), and ranges from around 1 year to the early 'teens' (Paradis, 2009). Whether invoking of learned rule substitutes for absent implicit understandings compromises the naturalness and spontaneity of language output during language use depends upon the learner's familiarity with those rules and the conditions under which communication occurs. Under circumstances conducive to rule selection and application learners may supply spoken and written English rivalling that of their English L1 peers (Paradis, 2009). Sub-optimal conditions, conversely, as when learners must contend with distractions, noise and interruptions, typically result in first language habits expressing themselves in L2 output, often betraying the speaker's mother tongue (Marinova-Todd, 2003; Möhring, 2001). Such L1 intrusion may prove short-lived (months perhaps), long-term, or even lifetime persisting (Paradis, 2009) depending upon the success of *retuning*. Many EAL students fail to pass as native-like despite decades of L2 exposure (Jedynak, 2009).

[34] This includes word-meaning associations in so far as they arise implicitly.

Factors predictive of acquisition difficulties include chronological age, the language competence of interest, and the time point beyond which acquiring that competence typically fails (Lenneberg, 1967; Marinova-Todd, 2003). Morgan and Kegl's (2006) sign language study found children less ably acquired syntax beyond the age of 6 years, though continued to make gains between the ages of 6 to 10 years. In their study of Grade 1 and 7 immersion pupils, Harley and Hart (1997) reported that analytical ability (i.e. learning) alone predicted grammatical competence gains among students in Grade 11 and beyond. In regard to phonetic sensitivity (see e.g. Mack, 2003), subconscious preferential responsiveness to L1 'sounds' develops from as early as age 6 months (see also Kuhl et al., 1992). By the age of one year children typically no longer react (or react only minimally) to phonetic elements absent in the mother tongue though they may learn to do so under conducive circumstances (Werker & Tees 1984).

Whether chronological age impacts upon a child's ability to construct the form/meaning associations at the heart of vocabulary gain depends upon how substantially implicit learning contributes to such gains at all, together with the critical age beyond which acquisition fails. While the strong form of the conscious learning hypothesis[35] now appears untenable (see e.g. Williams, 2005), the longstanding presumption that acquisition contributes only minimally to gaining word-meaning pairs remains firmly intact (Paradis, 2009), implicit gains only revealing themselves in tests of implicit knowledge (Paradis, 2009: see also Williams, 2005). Despite the word retrieval difficulties that emerge in 'old age,' conscious learning ability does not exhibit the same age associated decline observed of implicit language acquisition capacity (Bley-Vroman, 2009). Adults therefore may gain vocabulary throughout life just as they do other declarative understandings. How many novel word meanings 'older' students master depends upon factors underlying learning successes generally: i.e. capacity and inclination to learn when opportunities arise along with the aptitude to select appropriate strategies given the learning goal (Marinova-Todd, 2003). Post critical age L2 students may therefore develop

[35] This is the presumption that children only learn word meanings under the +attention +noticing condition.

vocabularies comparable to their native English speaking peers. Child EAL learners, too, if familiar with relevant strategies may display equivalent vocabulary development (Jamieson, 1976) to their native English classmates.

Despite ongoing research, many questions remain unanswered. Variation in critical ages remains little understood, as does the abruptness of acquisition loss for alternative language competencies. Faruji (2012) questions the neurological basis of critical-age arguments, while others (see e.g. Singleton, 2001) ask whether the 'construct' itself evokes quite the same understanding across academic disciplines. The uncertainties leave ample scope for disagreement. While EAL instructors tend to deny critical ages altogether, for neuroscientists the evidence for abrupt onset acquisition failure appears obvious and compelling (Paradis, 2009, p.109). Resolving such issues remains essential to developing a comprehensive understanding of vocabulary development.

2.5.3 Moderating Factor 3: The learning environment

The term learning environment refers to the *teaching strategies* that inform pedagogical practice within the institution the EAL child attends, along with the *social relationship* the child establishes with English L1 pupils and teachers from whom s/he receives L2 input (Genesee & Nicoladis, 2006, p.336). A teacher's *strategies* depend upon his/her attitudes towards those from other cultures, any associations presumed –reasonable or otherwise– between ethnicity and response to tuition, the difficulties s/he acknowledges arising from the child's L1 background and, not least, linkages the teacher supposes between cultural affinity and learning motivation (Anderson-Clark, Green, & Henley, 2008; Chiswick & Miller, 2004; Elhoweris, Mutua, Alsheikh, & Holloway, 2005). As a rule, EAL children progress most satisfactorily –in English, but also other curriculum subjects– when teachers' beliefs prove well-founded, course books exhibit cultural sensitivity, and when educators respond to potential threats to teacher/student rapport and trusting relationships (Scrivener, 2011). Poor attainment stemming from language programs founded upon inaccurate theoretical premises has a long-documented history (See e.g. Gibbons, 2002) as do unimpressive learning outcomes arising from teacher insensitivity to alternative cultural values.

EAL children benefit from mutually respectful relationships with English L1 classmates since those classmates provide examples of well-formed target language input from which language mastery derives. School environments in which *all* students feel respected, valued and socially at ease do not necessarily arise without active leadership, however, despite a well-intentioned management and staff. As EAL children progress through the school system their cultural perspectives evolve, redefining the dynamics of student/student interaction along with the composition of social groupings and relationships with peers. Depending upon the congruence of 'home' and 'adopted' nation values, the moral, religious and social norms of a host society may present unenviable choices, the student's response to which potentially carry significant social implications: Should, for example, one reject values that the 'family' holds dear? What of the new culture deserves respect? Should one attempt reconciliations between norms of the 'home' and 'received' cultures despite non-trivial differences that divide them? While some children go on to successfully assimilate with their English L1 peers and teachers, others withdraw and seek solace among those from the parental (or caregiver's) cultural background. How far assimilation proceeds affects both the quality and quantity of the target language EAL input, determining whether that input derives primarily from English mother tongue classmates or the interlanguage of possibly marginalized non-native speakers. Should assimilation fail, and deprived of good models of language and/or practice opportunities, L1 processing habits typically continue to express themselves in L2 production, becoming more resistant to remediation over time (Swain & Lapkin, 1986). Selinker (1975) described this outcome as "fossilization." The term nowadays refers to any entrenched and stable output deviations from native speaker norms.

That second language mastery requires rather more than a diet of 'quality' input became clear once linguists turned to evaluate the Canadian Government's immersion language programs (Harley, 1992; Swain & Lapkin, 2005). Dating from the 1960s, the programs targeted monolingual English L1 students at a time when proficiency in French had become a divisive social and political issue. Despite children receiving extensive L2 exposure the results of the initiative proved equivocal (see Siegel, 2010; Harley, 1992). Reviews dating from the 1990s reported teachers reluctant to correct errors, preferring to de-emphasize formal instruction in preference for flooding children with L2 input. Lyster (1990) described the typical enrollee as displaying strong listening comprehension skills but little control of grammar and pronunciation, often exhibiting a distinct immersion dialect (p.170). Harley (1992) reported students as fluent, but went on to note that L2 output typically scored poorly in tests of formal accuracy. Summarizing available literature, Paradis (2009) concluded that immersion outcomes proved generally inconsistent. For some children L2 exposure indeed ensured communicative proficiency and high end-state L2 competence. For a sizeable minority, however, language gains proved unimpressive and amounted to little other than mastering the “pidgin” of their peers (Paradis, 2009, p.119).

2.6 Learning durability

For teachers, the benefits from spaced learning opportunities mean little in pedagogical terms unless those gains persist over educationally meaningful time periods (e.g. months, terms or years) as opposed to mere minutes or hours. Inseparable from learning, then, stands the question of “durability” and the factors upon which memory persistence depends. The current section moves beyond the psychological processes behind word-meaning gains to examine the mechanics and theory of memory preservation. The discussion will play an integral part in building a comprehensive account of how durable word-meaning gains arise from RR sessions, an issue the study returns to in Section 2.8.

2.6.1 What is it that makes a memory durable? The depth of processing (DOP) hypothesis

The DOP hypothesis (Craik & Lockhart, 1972) associates memory durability with how deeply a learner analyses the stimulus from which the memory originates, deeper analyses presumed to supply more enduring memory traces than shallower. The hypothesis correctly predicted that cognitively demanding semantic analyses establish more secure memories than simpler orthographic analyses (e.g. a focus on letter arrangements), and that orthographic analyses, in turn, yield more secure memories than shallow acoustic analyses. Recalling its early 1970s origins, Wilson (2011, p.264) writes of the hypothesis' obvious "intuitive appeal." It deemphasized, for example, rehearsal in short-term memory as conditional for memory consolidation, such rehearsal having proved a contentious aspect of Atkinson and Shiffrin's (1968) memory model; it encouraged a potentially productive notion that envisaged memories less as configurations of neurological structures than the processing operations from which they derived. Not least, DOP provided both linguists and psychologists with a novel and intriguing new research perspective –a fruitful means to explore memory durability from outside of the conventional store-based contexts (Lockhart, Craik, & Jacoby, 1976).

Despite early signs of promise, as originally conceived, the hypothesis proved short-lived. 'Depth' resisted linguists' efforts at a definition (Craik & Tulving, 1975; Lockhart & Craik, 1990) and remained at odds with suggestions of qualitatively different processing operations (acoustic, semantic, orthographic). Revised models (see, e.g., Craik & Tulving, 1975) argued for independent processing domains, denying the original DOP claim that different forms of analysis (e.g. acoustic or semantic, for example) need occur in an invariant order (Craik, 2002). To depth, revisionists also added breadth of processing (i.e., elaboration) to acknowledge enrichment of encoding within hypothesized levels. By the late 1970s DOP proponents had come to conceive of depth and elaboration acting conjunctively to impart on a memory trace the quality of distinctiveness (Craik & Tulving, 1975). Distinctiveness now became the source of memory persistence, rather as distinct visual stimuli prove more discriminable in the visual field than those less so (Craik, 2002, p.307).

Even in its revised form, however, difficulties remained. Kolers and Ostry (1974) found repeated shallow operations supplied memory traces as durable as those from minimal semantic analyses. Craik (2002) cited failures to integrate stimulus-driven bottom-up processes and conceptually driven top-down processing into a workable DOP-based model (Craik, 2002; Craik & Tulving, 1975) while others objected to DOP as more descriptive than explanatory (Eysenck & Keane, 2005) or questioned whether the hypothesis satisfactorily explained incidental learning behavior (Roediger & Gallo, 2001). With the concerns still unresolved, by the late 1980s DOP persisted as little more than the proposition that more durable learning arises from greater cognitive effort (see, e.g., Bartlett, 1932; Smirnov, 1973). For this, at least, the evidence was, and still remains, compelling (Laufer & Hulstijn, 2001); indeed, it accounts for textbook writers' frequent reminders that teachers engage students in deep-processing tasks (see, e.g., Benjamin & Crow, 2010). The signs that DOP might have rather more significant implications for EAL instruction, including vocabulary learning, date from the early 2000s. The resurgence stemmed from the centrality of DOP to Laufer and Hulstijn's (2001) construct of involvement load (IL).

IL denotes the commitment an activity elicits and therefore the depth of processing it induces. Drawing upon theoretical and empirical studies, Laufer and Hulstijn (2001) argued that the greater a task's IL aggregate score, the more learning that arises during task completion; this they termed the involvement load hypothesis. An IL score equals the sum of three independent factors: *need, search,* and *evaluation.* Each factor is presumed either absent or present and, if present, exists in a moderate or strong form. The hypothesis considers strong forms as more learning conducive than moderate, and moderate forms more conducive than absent forms. The elements of IL are as follows:

Need: This represents a motivational, or non-cognitive, aspect of involvement. The term denotes a 'desire to achieve' in the sense of willingness to comply with task requirements, as opposed to "failure avoidance" (McClelland, Atkinson, Clark, & Lowell, 1953). Moderate need exists when the motivation for task completion stems from an external agent such as a class teacher who might require, for example, that students complete a particular exercise or test. Strong need, conversely, originates from student internal needs and aims –a personal drive towards achieving the task goal. Typical behaviors associated with need (moderate or strong) include looking up a word in a dictionary or thesaurus during essay writing or seeking clarification of word meaning from a teacher during a reading activity (Laufer & Hulstijn, 2001).

Search: Search represents one of two cognitive elements of involvement, the other being evaluation. The term refers to the effort a learner expends to assign meaning to a word "or … to find the L2 word form expressing a concept" (Laufer & Hulstijn, 2001, p.543). Like N*eed*, typical search behaviors include consulting a teacher or referring to dictionaries or thesauri. A zero search task calls for no meaning-derivation effort. Laufer and Hulstijn (2001) cite the example of a reading assignment in which translations of unfamiliar words appear in glosses or marginal notes.

Evaluation: This refers to comparing the meaning of one word with another, or the particular meaning of a word with any alternatives the context might support (Laufer & Hulstijn, 2001, p.544; Laufer & Girsai, 2008). Evaluation amounts to moderate when the learner attempts to discern differences or similarities in meanings between two words as, for example, when completing a fill-in task with candidate words provided. Strong evaluation involves decisions regarding how "additional words will combine with new words in an original sentence or text" (Laufer & Hulstijn, 2001, p.544).

Empirical tests of the ILH begin with Hulstijn's (2001) parallel study of incidental acquisition of ten target words by young EAL students in two locations, Israel (n=99) and the Netherlands (n=87). At each site participants formed three groups (random assignment), each of which completed one of three tasks differentiated by IL rating. For Task 1 (*reading task*; low IL= 1= [+N (Need), -S (Search), -E (Evaluation)]) participants read a short text with novel vocabulary highlighted in bold and 'glossed' in the margin. A post-reading test consisted of text-related comprehension questions. Task 2 (*fill in*; moderate IL=2=[+N, -S, +E]) employed the same text but with target words (n=10) deleted, study participants (n=10) having to infer the meanings and match these with supplied translation equivalents. For Task 3, each test taker wrote a letter that included specified terms, definitions of which appeared on an accompanying sheet. For this Hulstijn (2001) assigned an IL rating of 3 (i.e., no search, moderate need, and strong evaluation [+N, -S, +E]. All post-tests took place on two occasions: immediately after completion and then again one week (Netherlands) or two weeks (Israel) later.

Hulstijn (2001) describes the findings (results from a 3 X 2 ANOVA) as broadly ILH supportive noting that students gained significantly more word meanings from Task 3 (the high IL-rated writing assignment) than either 1 or 2. Results from a Newman–Keuls analysis of the Dutch data denied a more robust affirmation, the findings revealing similar mean scores from the ‘fill-in’ and ‘read-only’ tasks despite IL ratings of 2 and 1 respectively. Whether this counterintuitive result stemmed from a fault with the hypothesis or the research methodology remained unclear. Applying the same test to the Hebrew-English data, however, supplied results fully consistent with IL predictions.

More recently, Martinez-Fernandez (2008) reexamined the IL hypothesis employing a think-aloud protocol that required participants (college students, n=45 enrolled in a fourth-semester Spanish course) to verbalize their thoughts while completing assignments each differentiated by IL scores. The author based the assignments upon *Shade for Sale: A Chinese Tale* (Dresser, 1994), the text duly adapted to include select target words. *Assignment 1* consisted of a comprehension exercise (low IL=+N, -S, -E) in which glosses of each target word appeared at relevant points in the script. For *assignment 2*, the reader viewed three alternative glosses for each target word from which s/he then selected the most appropriate (high IL=+N, +S, +E). *Assignment 3* (intermediate IL=+N, -S, +E) required participants to provide the missing target words in a short text with several definitions available from which to make a selection. The study design involved recording and then analyzing the written transcriptions of participants’ verbalizations while engaged in task completion. From this data source, Martinez-Fernandez (2008) went on to derive awareness scores (i.e., degree of understanding) for three types of language features: (a.) ‘word form’ only, (b.) ‘meaning’ only, or (c.) ‘word and meaning.’

The findings proved equivocal, but generally unsupportive of ILH predictions with the low IL fill-in task inducing “significantly more awareness” (Martinez-Fernandez, 2008, p.225) than all others including the high IL-rated multiple-choice assignment. Participants’ scores revealed neither deeper processing, nor more vocabulary development, from higher-rated tasks with the latter failing to prompt “deeper processing” (i.e. high awareness) at either assessment occasion (p.227).[36] The study

[36] Participants undertook assessments immediately after task completion and then again one week later.

acknowledged, however, several methodological concerns that potentially explained the counterintuitive findings: for example, the possibility that concurrent verbalization might moderate the attention a task elicits (Leow & Morgan-Short, 2004), that test takers necessarily fail to articulate subconscious contributions to task completion behavior, and that 'task type' might influence the manner of task performance (Jourdenais, 1998; Yoshida, 2007). More pragmatically, the possibility remained that participants may, on occasion, have simply neglected to communicate relevant thoughts or unsuccessfully conveyed their thought processes. To the extent such omissions or communicative failures arose, genuine instances of awareness went unrecorded.

Other attempts at ILH corroboration have proved supportive, partially so, or tending towards outright rejection. Huang (2004) found that IL (multiple-choice, gap fill, or sentence making) successfully predicted vocabulary uptake among her (Nanjing) university-level participants. Kim's (2008) adult participants displayed more initial, but also delayed, learning from high-IL (vocabulary-based) tasks than those assigned lesser ratings. Keating (2008) claimed that IL accounted for differences in vocabulary uptake among "beginning learners of Spanish" though noted lesser benefits after controlling for time on task. Moonen, Graff, Westhoff, and Admiraal (2005, p.43) likewise reported favorably of the ILH, describing the hypothesis as a valuable contribution to understanding incidental vocabulary acquisition, though argued for clearer distinctions between strong and moderate forms of IL factors. Why, for example, would 'strong' evaluation contribute more to task involvement than 'moderate' that requires comparing different meanings of the same word? Cheng (2011) likewise asks for terminological clarification, though reported "partial" (p.84) support for the ILH from the vocabulary retention scores of 111 tertiary-level Taiwanese participant volunteers having controlled for trait anxiety. Yaqubi, Rayati, and Allemzade (2010) agree with Moonen et al. (2005) that properties of the task itself may contribute to learning irrespective of IL rating. If correct, the predictive power of the ILH depends upon both IL score along with the behaviors a task requires for completion. EAL teachers have tended to raise more practical concerns centering around the ILH's relevance to child learning given the preponderance of studies involving adult participants. Not least among the unresolved issues remains the question of how need, search and evaluation might contribute to learning outcomes independently of one another.

2.7 The question of data elicitation

Research claims remain only as valid as the data-elicitation process and methodology from which they derive. From poor data comes untenable conclusions that subsequently reveal themselves in studies employing more robust experimental designs. While concerns in regard to incidental word-learning research have long attracted academic commentary (see Read, 2000 for a review), one issue stands out above all: the threat to robust data collection from multiple-choice test instruments (MCTs). Apart from a raft of general, and oft-cited, objections to this test format (see Anderson & Freebody, 1981; Meara & Buxton, 1987; Wesche & Paribakht, 1996), Horst (2001) cites two issues of particular relevance to the vocabulary researcher: (1.) the MCT's ill-suitedness to detect partial knowledge gains, and (2.) that such tests provide poor indications of whether a child knows a word to the criterion standard. Failure to capture partial knowledge means MCTs typically lack the sensitivity for exploring the incremental, piecemeal, manner in which vocabulary development proceeds. The child who fails to select the correct 'answer' from a choice of distracters may indeed have no familiarity with the target word or, and quite plausibly, only marginally less than suffices to have answered successfully. To identify intermediate lexical knowledge between the polar values of 'known' and 'unknown' typically calls for several MCTs each differentiated by difficulty ('easy,' 'average,' and 'hard'), the teacher directing a student who performs impressively on an 'easy' test version to 'move on' and attempt a more challenging variant. Nagy et al. (1985) employed just such multi test assessments in their several studies of incidental vocabulary uptake among English L1 high school children (see Figures 2.1 and 2.2). Such tiered-tests have, however, proved time consuming to design and validate leaving an ongoing need for more practical test formats. The most popular of the potential alternatives make use of self-assessment and 'knowledge states. Examples of instruments adopting this approach include Waring's (2000) State Rating Task, and Paribakht and Wesche's (1993) Vocabulary Knowledge Scale (see chapter 3).

Level 1: Gendarme means:

a. to trick or trap someone
b policeman
c. spoken as if someone were out of breath or having difficulty breathing.
d. the secret collection of information about another country
e. the illegal transportation of goods across a border
f. I don't know.

Figure 2.1: Multiple-choice test (easy), from Nagy, Anderson and Herman (1985, p.240).

Level 3: Gendarme means:

a. policeman
b bellboy
c. bodyguard.
d. spy
e. waiter
f. I don't know.

Figure 2.2: Multiple-choice test (difficult), from Nagy Anderson and Herman (1985, p.240).

Horst's (2001) second concern with MCTs -the issue of 'criterion standard'- arises from controversies over MCT design. Despite an extensive literature addressing test construction, disagreement persists over the optimum number of distracters, the merits of distracters in the test taker's L1 (Read, 2000), whether tests should contain penalties for guessing and the utility of the now commonplace '*I don't know*' option. This absence of consensus complicates cross-comparisons between studies. While Nagy, Herman, and Anderson (1985) presented their test takers with six distracters for each target word, Day, Omura, and Hiramatsu's (1991) participants selected from just five, implying performance on one MCT may poorly predict scores on an alternative assessing the same word corpus. Even among tests constructed upon shared notions of appropriate wording, number of distracters and common conceptions regarding L1 use etc., researchers may yet disagree over the competencies designative of ‘word knowing,’ or their relative weightings in determining a test-taker’s lexical knowledge. At times, debate arises over more fundamental issues still. Should ‘play’ and ‘playing,’ for example, count as one word or two? Should we analyze an idiomatic expression as a word or phrase? And just how reliably do receptive knowledge tests indicate productive knowledge skills? As Read (2000) explains, a child may know a word to the satisfaction of one investigator but not another (Read, 2000).

A longstanding issues among MCT designers concerns how best to correct for student guessing. A common correction formula dates from the 1920s and remains a popular subject of academic discussion:

$S = c - (E / n - 1)$ Where:

S = the corrected score
E = the number of incorrect responses
c = sum of correct responses
n = number of options

The formula’s limitations find a place in most standard texts dealing with test design (see e.g. Haladyna, 2004). Apart from potentially negative corrected scores, the formula may underestimate true attainment –the more so the fewer the distracters available for selection. Given a 50-item test, four distracters and a ‘true’ score of 38 correct responses (i.e., the learner indeed knew the meaning of 38 words) the amended score amounts to 34, a reduction of 4 points. Had the test taker known the answer to 25 questions, the amended figure falls to a notably less impressive 16.7. The search after meaningful scores becomes more challenging still should a test taker have sufficient knowledge to reject one or more distracters from the options available. In Day, Omura, and Hiramatsu’s (1991) MCT, the child who successfully discounted one distracter of the five for each target word raised the likelihood of a correct guess from 20% to 25%. Whether including an '*I don't know*' distracter discourages such 'elimination' behavior depends upon the test taker's aversion to risk and/or his/her understanding of the penalties incorrect choices entail. Roberts (2006) found that students who willfully disregarded the option often achieved higher scores than those who did not. The most able multiple-choice performers, it appears, draw upon two knowledge types (Merritt, 2006): (1.) an understanding of the relevant subject matter (i.e. what a test claims to measure) and, (2.) technical familiarity with MCT as a test genre. The highest scoring MCT takers, in this view, typically include those who understand and exploit the principles informing MCT construction. As Merritt (2006) explains, "Research shows that students can perform significantly better on objective tests by improving their test-taking skills" (p.10).

For the conventional 'matching' type MCTs, Nation (2000) stresses the importance of intelligible distracters. He calls for simple syntax (Compare "John built the ..." with "The … was built by John.") and ensuring that the vocabulary of distractor choices consists of more commonly occurring words –and hence more likely known words– than the tested item.[37] Nation's (2000) presumed correlation between frequency of occurrence and likely familiarity with a word's meaning has indeed proven "robust" (Milton, 2009, "26). How dissimilar in 'occurrence frequency' the distracter and target word need be, however, remains unclear. To test a word among the 8,000-9,000 most commonly occurring, should the researcher, for example, build distracters from vocabulary within the 1-6,000 frequency band, the 1-7,000, or 1-8000 or some other range? The issue raises complications when designing tests for children with small vocabularies, or post-critical-age learners having only recently embarked upon second language study. As a broad generalization, the frequency/familiarity correlation can, moreover, become misleading if applied to individual students or discrete populations (Milton, 2009). 'Homework' and 'fork' both rate as relatively rare in the British National Corpus (BNC), appearing as the 6,327th and 6,085th most common words respectively and yet will likely prove

[37] Various word lists are available, the most notable being perhaps the British National Corpus (BNC) (see Aston & Burnard, 1998).

familiar to a 5-year-old schoolchild with a vocabulary of 4,000–5,000 word families (Schmitt, 2000, p.3). For learners familiar with a 1,000 or so L2 words, designing intelligible and natural-like distracters in the target language may prove impractical or impossible (Read, 2000). The concern prompted Read (2000) to propose tests that would permit a child to match L2 terms with his or her first language translation equivalents. The place of the mother tongue in language testing, however, remains contentious (Liu, 2009; Mattioli, 2004).

2.8 Putting the pieces together

Previous sections of this chapter have discussed the role of attention in learning, the forms of learning from which vocabulary gains arise and the several factors affecting the durability of gains over the short and long term. The current section integrates these findings into a plausible account of how children gain new word meanings during an RR session. This discussion serves two ends: (1.) It identifies, and clarifies, the role of factors that contribute to learning outcomes during RR; this will inform the methodology of the current study, and (2.) It locates the particular contribution of spaced learning within the context of the totality of factors that explain word meaning gains from reading experiences.

Section 2.5.1 argued that word-meaning gains from reading arise from several distinct behaviors. For emergent readers, as opposed to their more accomplished peers (i.e. lexical readers), the first step involves decoding and blending orthographic characters. This serves to recover a word's phonological form (Adams, 1990). Once the form becomes available, the reader then engages in meaning-derivation,[38] perhaps employing 'guessing from context' (GFC; see below) or consulting reference sources –a dictionary or thesaurus, for example. Finally, s/he consolidates newly gained form-meaning associations in memory, possibly through explicit learning, allowing for subsequent receptive or productive use (Bahrick, 1984; Schmitt, 1997). Typical consolidation strategies include

[38] Unless the meaning is obvious from the text (see non-explicit incidental learning above).

use of ‘word cards’ or software but could also arise purely from repeated instances of *noticing*. For accomplished, lexical (p. 41), readers decoding typically proves unnecessary since familiar (often-read) textual vocabulary will evoke a near instant match with a stored meaning in long-term memory. Only when encountering novel words do such proficient readers resort to a conscious decoding effort. The common behavior among capable and emergent readers is GFC –arguably the most ubiquitous and effective vocabulary learning strategy both adults and children apply during recreational reading sessions (Nation, 2001).

The term GFC refers to two strategies: (1.) *formal guessing,* and (2.) *contextual guessing*. Formal guessing refers to deriving word-meanings from attending to affixes, roots, word stems and morpheme knowledge. It describes how those familiar with Latin or Greek might attempt to decipher *thermoluminescence* (Read, 2000), or how a child familiar with English affixes might assign meaning to *un*happy, *im*possible or *in*correct. The second strategy, contextual guessing, refers to deducing meaning from clues within larger structures such as clauses and paragraphs, drawing upon any prior understanding the reader may have of the text’s theme or subject matter (Diakidoy, 1998; Paribakht & Wesche, 1999; Pulido, 2004). A keen footballer, for example, might deduce the meaning of ‘penalty’ from a match report, or a chess player, the meaning of ‘check’ from an account of a game. Which GFC a reader employs depends upon his/her learning style, attributes of the word itself (whether it displays morphological transparency or opaqueness), textual factors such as the helpfulness of contextual clues (Ames, 1966), background topic knowledge, and whether the strategy proves accessible to the child at all (Beck, McKeown, & McCaslin, 1983). A child cannot formally guess if unable to partition words into familiar morphological units. Nor can s/he make use of contextual guessing should reading comprehension prove insufficient for word meaning inferences. As a general rule, the more effort expended in GFC, the more durable the child’s learning outcome (see the Involvement Load Hypothesis, p. 60).

How well a child deduces word meanings depends upon the aptitudes s/he applies to what amounts to a conscious, deductive, learning task (Paradis, 2009; Bloom 2000). These aptitudes consist of genetic attributes such as verbal IQ and working memory capacity, learned skills such as GFC (among others) or composites of both genetic and learned competencies such as the ability to select and implement meaning derivation strategies. The "learning burden" (Nation, 2001, p.23) children encounter during meaning deduction depends much upon the structural similarity between the target language and mother tongue or, as Chiswick and Miller (2004) explain, the language distance between them (Section 2.5.2). A list of possible lexical dissimilarities appears in Swan (2006):

a. Difficulties learners experience with words such as "shame, remorse, apology, repentance, and penance" (Swan, 1997, p.159) given distinct, culturally laden connotations;

b. Difficulties arising from culturally bestowed notions of "word" itself; speakers of synthetic languages (e.g., Finnish), for example, will attribute "more semantic information to words than do speakers of, say, English or Swedish" (Ringbom, 1978, 1986, 1987 p.155).

c. Difficulties arising when words in related languages have similar meanings but vary in their permissible collocations or grammatical structures. The French *expliquer,* for example, unlike the English *explain* may take an indirect object without a preposition (Swan, 2006).

d. Difficulties that arise in the absence of unambiguous translation equivalents. Swan (2006) cites the learning burden Japanese native speakers experience with English adjectives arising, in part, because the corresponding Japanese terms function as nouns.

How a child might best respond to the learning burden depends upon (a.) the lexical competence in which s/he proves deficient (i.e. just what s/he has yet to master of the L2 vocabulary) and (b.) the manner in which the competence naturally arises among English mother tongue users (see Section 2.2). Should the competence develop implicitly,[39] then the EAL learner has two options: either sensitize the implicit system (see Section 2.5.2) thereby allowing L2 gains as they would arise during L1 development or, second, employ learned rule substitutes for the absent implicit understanding. Each response calls for *noticing* behavior. For sensitization, *noticing* (response 1) raises a language detail to consciousness thereby allowing the child to construct an SR to which the implicit system will subsequently respond by tallying elements of L2 intake it now 'acknowledges' by virtue of prior *noticing* events (Section 2.2); for building conscious understandings (response 2), *noticing* alerts the child to language details s/he may then 'deliberately' access to build pedagogical rules allowing him/her to generate formally correct language output (Harley & Wang, 1997; Paradis, 2009). Among post critical age students, in particular, learning provides a valuable route to building a stock of rules able to serve as substitutes for absent implicit understandings. Minimally, this learning effort entails *noticing* relevant surface structure forms or, and rather more productively, will involve deliberate language study. Since maturation compromises learning less so than acquisition (Lenneberg, 1967), EAL students will continue to gain both word meanings and pedagogical grammar rules beyond the age(s) at which subconscious L2 mastery fails (Section 2.5.2). A child deprived of language exposure until his/her 'teens' (feral children such as Genie, Kaspar Hauser, and Isabelle, stand as examples[40]) will fail to acquire competencies that native speakers gain implicitly during early life, the implicit system by now having become relatively inaccessible (see Denham & Lobeck, 2009). The same child may, however, go on to learn form-meaning associations and display impressive vocabulary development.

Should the L1 and L2 possess words that denote a notion common to both languages, a child need not relearn that notion if already familiar from L1 vocabulary development experiences; rather, s/he assigns to the novel L2 term the meaning of the presumed first language translation equivalent

---

[39] That is, were it not for the learner's age and loss of implicit system plasticity.
[40] See, e.g., Newton (2002) for details.

(Section 2.5.2). Subsequent opportunities to *notice* the word in informative contexts will assist him/her in further disambiguating the L1 and L2 terms through a process Stringer (2008) terms relexification. How cognitively demanding relexification proves (the process calls for a conscious learning effort) depends upon the similarity in meaning between the percepts to which the respective first and second language words refer. For near synonyms, learning target language meanings typically proceeds with relative ease. For only partially synonymous words, however, relexification may call for prolonged and deliberate study.

To the 'relexification challenge,' EAL students confront further difficulties stemming from properties of words themselves and/or the texts in which they appear. Elley (1989), Huttenlocher (1974) and Gentner (1982), for example, reported associations of learning ease with lexical class. Gleitman (1990) noted children's particular difficulty acquiring verb referents, while Schwanenflugel (1991) related learnability to the property of concreteness –how imageable a word proves in terms of visual or acoustic distinctiveness (compare 'pencil,' an object imbued with shape and color, with 'courage' possessing neither). 'Word length' predicted word gains in Nagy et al.'s (1987) reading investigation involving high school students with shorter words as measured in number of letters proving somewhat more learnable than longer. Other studies have reported comprehension difficulties arising from excessive unknown word density. Hu and Nation (2000), for example, claimed 'inadequate' comprehension among pre-university volunteers (see also Laufer & Sim, 1985) once 20% of words in academic text proved unfamiliar. At 2%, comprehension rose to "satisfactory" but then fell to "minimally acceptable" when unknown word density stood at 5%. Comparable figures may apply to children. Drawing upon pupils' self-reports, Carver (1994) concluded that primary school reading materials should ideally expose a child to no more than five unknown words in every hundred.

Other factors impacting upon vocabulary gains include text genre and reading difficulty. Rice (1986) found that adults with extensive vocabularies tended to "read more sophisticated materials" (p.102) such as technical journals and science magazines. Cunningham (2005) claimed children gain more words from narratives than expositories, while Gardner (2004) reported readers who would assign alternative meanings to the same novel word depending upon the text type they engaged with. The readers' learning behavior (see section 2.3) -and how much new word learning occurs during RR sessions- may also depend upon a sense of 'study worthiness.' Gardener (2004) suggests children deem certain texts, e.g. comics, as typically less deserving of serious attention than others such as works of non-fiction.[41] The child encountering an unknown word in '*The Beano*' would therefore dwell upon the meaning rather less so than had the word appeared in a script explicitly seeking to 'explain' or 'teach' (Gardner, 2004). Shokouhi and Maniati (2009) add that structural differences between genres moderate word gains through their impact on comprehensibility and the utility of *guessing from context*. They single out narratives as "less cohesively organized by temporal and causal connections…" and demanding of "more explicit logical inference" (2009, p.15).

Whether readers encounter novel words under massed or spaced conditions, the likelihood of word gain depends much on comprehension difficulties associated with the complexity of the subject matter, the syntax, and the author's writing style. Since vocabulary learning arises from deductive efforts, readers gain fewer words from overly challenging texts than simpler alternatives requiring less diversion of cognitive effort from word meaning derivation to resolving non-lexically induced comprehension challenges. The more such diversion, the less pleasure a child typically derives from a script, the slower his/her reading rate becomes and the less vocabulary s/he gains per unit of time. The 'ideal' RR text should prove 'difficult' but not unduly so (Krashen, 1988) – a text, that is, that satisfies the test of i + 1 (Section 2.5.1) and succeeds in drawing the child to the border of his/her reading competence. The factors most predictive of the breadth and depth of a child's word gains over the long term include the proportion of texts s/he reads of particular genres, the complexity of those texts, and the forms of conscious learning the reading material the child chooses to engage in (Section 2.3).

Learners consolidate form-meaning associations either from (a.) strategies they apply for this purpose, (b.) repeated instances of *noticing* the same novel word, or (c.) the encoding operations that

[41] For example a child who reads non-fiction will likely engage in explicit learning rather more than one who only reads fiction.

accompany successful meaning recall (Karpicke & Roediger, 2007; Thomson, Wenger, & Bartling, 1978). The optimum interval between encounters with the same novel word to minimize learning time varies from reader to reader and depends upon the circumstances under which reading occurs (whether for example, reading to learn academic content or purely for pleasure), metacognitive skills and how advanced is the child's *Theory of Mind* (Section 2.5.1). For establishing word-meaning associations Bahrick (1984) described the ideal interval as sufficiently 'long' that a learner would barely recall the novel word at the time of its next encounter. These intervals, initially 'short,' become progressively larger as a word meaning becomes more secure and memory decay less pronounced. How many encounters suffice to consolidate a form/meaning association in long term memory depends upon the frequency and length of the intervals (Section 2.4) together with factors predictive of learning success generally e.g. the child's familiarity with word-learning strategies (see p.64) and his/her cognitive maturity. Kachroo (1962) found that schoolchildren could learn novel words in textbooks if those words recurred seven or more times but only rarely so if encountered just once or twice. Nagy et al. (1985) reported that 10 to 20 encounters with the same word sufficed for a child to identify the correct distractor on an MCT vocabulary test while Swanborn and deGlopper (1999) suggested a 15% probability of word learning from a single textual encounter though noted variability associated with the learner's affective state, chronological age, and metacognitive skills.

2.9 Summary

This chapter began by defining a distinction between conscious *learning* and subconscious *acquisition*, before going on identify the particular lexical competencies associated with each process. The chapter argued that word-meaning gains arise from a conscious learning effort where consciousness implies, minimally, the act of *noticing* an association between a meaning and an orthographic or phonological form. Building upon findings from spaced learning research the chapter went on to argue that the time intervals between reencounters with same novel word during reading experiences have an important bearing on whether a child establishes a form/meaning association and the likelihood s/he consolidates this in the long-term memory store. During the course of the discussion, the chapter identified moderating factors that qualify the efficacy of spaced learning during RR sessions –factors, for which controls will prove necessary in the main study (see Chapter 4). A concluding section pulled together findings into a comprehensive account of word learning from reading experiences. This latter section defined the place of spaced learning within the broader context of the many determinants of successful learning outcomes and suggested a role for the spacing effect in a robust account of vocabulary development from RR sessions.

Chapter 3

Methodology 1: The data elicitation instrument

3.1 Introduction

The previous chapter reviewed research findings that shed light on how word learning arises from RR experiences before then moving on to synthesize those findings into a plausible account of 'word-learning' within the context of school-based RR sessions. The current chapter addresses the altogether more pragmatic issue of how to gain reliable and robust indications of vocabulary gain from RR sessions, a concern first touched upon in Section 2.7 (above). The chapter begins with a review of the VKS (Paribakht & Wesche, 1993, 1996, 1997), one of the more widely employed alternatives to the MCT test format (Bruton, 2009). Drawing upon that discussion, the chapter moves on to describe the design of an alternative instrument, the VSAT,[42] for use in the present investigation. The remaining sections of the chapter report the results of three pilot studies designed to establish VSAT suitability for classroom use in the host institution.

3.2 The Vocabulary Knowledge Scale (VKS)

The VKS evolved as a response to concerns regarding MCTs for vocabulary-assessment purposes (for a discussion, see Horst, 2001; Wesche & Paribakht, 1996). The instrument recognizes five stages of vocabulary growth, each stage representing a step along a presumed lexical developmental path extending from total unfamiliarity with a target word to the rich understandings of a hypothetical literate native speaker. The stages intentionally correspond to the developmental levels of Gass's (1988) Integrated Model of L2 acquisition:

[42] Designed by the researcher, this is essentially a modified version of the VKS.

1. **Apperceived input** (for a discussion, see '*noticing*,' Section 2.3): "…that bit of language that is *noticed* in some way by the learner because of some particular features" (Gass & Selinker, 2001, p.400).

2. **Comprehended input**: input that is comprehensible and comprehended.

3. **Intake**: "The process of assimilating linguistic material" –see Gass, (1997, p.5) for a discussion.

4. **Integration**: The internalization of the new word meanings.

5. **Output**: The application of word knowledge meaning for productive and/ or receptive use.

A typical VKS session begins with the researcher presenting pre-prepared target words (usually on cards) to the participant, one at a time, pausing between presentations as the test taker assigns each word to a state, either by placing the word card on a mat displaying the VKS descriptors or indicating orally to which state s/he believes the word belongs. Following the assignment, testing proceeds according to the word's state placement. Should assignment be to States 1 or 2 the administrator promptly moves on to present the next target word from the list, the test taker's self-assessment presumed a reliable indication of lexical competence. Assignment to VKS States 3, 4, or 5, however, requires verification of placement accuracy (Figure 3.1). This involves two steps: first, the administrator elicits from the test taker a response to a prescribed verification question; second, s/he assesses that response against the appropriate interpretive criteria (Figure 3.2). Depending upon the assessment outcome the word either remains in the state to which initially allocated or the administrator assigns it to another, deemed more representative of the test taker's lexical knowledge. This reassignment process may itself prompt a further round of questioning and word movement; the arrows in Figure 3.2 indicate this possibility. The final state a word occupies determines its numerical score, a figure falling between 1 and 5, with 5 denoting the highest 'word' level of understanding and 1 the least. Having now assigned each word a score, the researcher (optionally) then proceeds to calculate an average value that serves as a general measure of test taker familiarity with the target word corpus.

The Vocabulary Knowledge Scale (VKS)

1. I *don't remember* having seen this word before.
2. 2. I *have seen this* word before, but *I don't know* what it means.
3. I *have seen* this word before, and *I think* it means … (synonym).
4. I *know* this word. It means … (synonym or translation).
5. 5. I *can use* this word in a sentence. (participant asked to provide a sentence).

Figure 3.1: The Vocabulary Knowledge Scale (VKS) (Paribakht & Wesche, 1993).

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

Figure 3.2: The VKS scoring procedure (adapted from Paribakht & Wesche, 1997).

For tracking and measuring vocabulary development the VKS remains highly regarded. Bruton

(2009) describes the VKS as ‘popular’ for data-elicitation citing its use in studies of adult lexical progression. The instrument requires little technical expertise of administrators, boasts a track record of yielding useful data from tertiary-level students, and derives credibility from exhaustive testing of internal and external validity. In *Second Language Vocabulary Assessment*, Read (2000) describes the VKS as “practical” for exploring a word’s initial recognition and use. Nation (2003) noted the instrument had gained “significant currency in second language vocabulary assessment,” while Horst (2001) calls the VKS a principled response to the need to measure vocabulary depth. Bruton (2009), albeit citing several objections to aspects of VKS design, acknowledged the instrument as “probably” the “best known measure for assessing productive and receptive vocabulary from L2 reading” (p.288) citing its common place in research since the mid-1990s (e.g., Joe, 1995, 1998; Pulido, 2004; Rott & Williams, 2003).

Objections to the VKS broadly fall broadly into two categories (Bruton, 2009): (1.) complications arising from applying arithmetical operations to nominal data (Waring, 2000) and (2.) practical issues stemming from efforts to verify word placement. The first category encompasses statistical concerns associated with summing and averaging figurative (categorical) labels for VKS descriptors. As qualitative, ‘literary,’ statements, Waring (1999) stresses that scores amount to impermissible ‘input’ to arithmetical expressions calling for quantitative data. In no meaningful sense, that is, can one speak of an average competence between *I can use this word in a sentence* (Score 5) and *I have seen this word before* (Score 3). Nor could any such purported average carry an associated standard deviation or other indication of dispersion about the ‘mean.’ An average of 4 will say nothing of how many words might occupy State 3 or State 5, or even whether the test taker knew any words to the State 4 standard. The researcher gains no definitive indication, in short, of just how many words from a corpus lie within a VKS state of interest. Waring and Nation (2004, p.11) cite the illustrative case of two hypothetical students with average scores of 3:

A: 1111155555 = 30/10 (average 3)

B: 2414232534 = 30/10 (average 3)

Associations between VKS ‘score’ and ‘lexical competence’ become more tenuous still given the qualitative distinction between the productive understanding State 5 seeks to capture and the receptive skills associated with States 2 and 3. This difference in knowledge ‘kind’ means few linguists accept the implicit presumption one can conflate different competencies into a single construct (Read, 2000), preferring instead to assign each competence its own measurement scale. Indeed, this very objection to ‘word knowing’ as a unitary, indivisible, notion prompted Henriksen (1999) to propose a multi-continua depiction of word knowledge (Section 1.4), and Waring (2000) to argue for measurement instruments that posit functionally independent knowledge states. While the VKS arguably captures a lexical knowledge hierarchy of sorts (a language user, after all, has more knowledge of a word occupying State 5 than State 3) the relationship between ‘scores’ and ‘states’ bears little correspondence to the more authentic representation of continuous data (e.g. temperature or water pressure) on interval scales. A child does not ‘know’ a word occupying State 4 twice as well as s/he does another occupying State 2, or four times as much as a word in State 1. Even within a VKS state, words may not prove equally familiar since some will better exemplify the relevant descriptor than others. A word in State 4 may lie on the border of either State 5 or State 3 but, and just as plausibly, occupy any point between. Adding additional states –nothing says an instrument cannot incorporate more than 5– will reveal more precise indications of lexical understanding, as well as address criticism (e.g. Meara, 1996) that the range of competence a state designates appears excessive. This expedient, however, calls for complex, elaborate, and time-consuming verification procedures given the more discriminating word knowledge assessments test takers must now engage in.

The second category of concern –the practical difficulties of verification– arise when test takers assign a word to State 5. It remains critically important that researchers reliably identify pre-learned clauses (rote learned sentences in textbooks, or examples of usage in dictionaries) more sophisticated than the test taker could reasonably generate (Hakuta, 1974). With no opportunities to uncover such language during testing (time rarely suffices for such ‘probing’), VKS administrators must rely on familiarity with learners to discriminate between memorized clauses and those that test takers authentically self-construct. Without that familiarity, testing inevitably inflates State 5 sums to supply exaggerated indications of lexical competence. The follow-up synonym test following a State 5 word placement raises yet a further issue –the question of appropriate language for student responses. Researchers who insist on L2 synonyms will potentially induce failure in a task possibly manageable were L1 responses deemed acceptable. Read (2000) argued that this supports mother tongue use in vocabulary testing, as has Liu (2009, p.65) who sees the L1 as helpful "for checking and validating L2 learners' understanding" (see also Atkinson, 1987). The merits of L1 usage for lexical testing remain controversial, however. Mattioli (2004) cautioned against first language intrusion in classroom settings while Tang (2002), though sympathetic to L1 references, acknowledged "commonsense assumptions" arguing for total L2 immersion. Among EAL teachers themselves, a preference for 'English only' lessons seems long-established, Howatt (1984, p.281) describing school policies prohibiting the first language as the very "bedrock" upon which other principles of L2 teaching derive. A decade later and little had changed, Scrivener (1994), for example, listing student L1 use as among the long-standing "problems" (p.192) confronting the E.A.L. instructor. The debate over mother tongue usage continues in school staffrooms at both home and abroad, often informing language regulations. While some schools might prohibit the L1 altogether, others lean towards permitting it under limited circumstances –typically in the playground but not the classroom.

Given the VKS's origins as an instrument for exploring tertiary students' vocabulary development, its suitability for use with primary aged children remains unclear. Butler (1990) reported only a poor correlation (r=0.38) between self-evaluations of 7-year-olds and teacher measures of language competence, while Okanda and Itakura (2007, 2008) claimed young children (at least up to age 5, but possibly beyond) display an affirmation bias in response to yes/no questions. Waring (2000) draws attention to a further concern: the particular interpretive difficulties a child faces given the choice of VKS test wording. Words such as "know" (State 4) and "use" (State 5) may not carry quite the same connotations among nine-year-olds as they do among adults, just as adults attach dissimilar meanings depending upon prior education and life experiences. Durso and Shore (1991) for example reported several instances of university graduates rating words as “unknown” only for the researcher to subsequently detect a measure of understanding. But even if researchers and test takers *do* agree upon definitions and ‘key wording’ the question remains of why a word occupies the state it does. The rationale for assigning a word to state ‘y’ as opposed to ‘x’ has little relevance for studies seeking only to quantify vocabulary breadth or depth; however, the issue becomes critically important for attempts to relate vocabulary development to a specific textual factor or factors. For such purposes, the VKS in its present form may have little to recommend it.

3.3 A new instrument: The Vocabulary State Assignment Task (VSAT)

The VSAT data-elicitation instrument represents a response to the difficulties that argue against VKS use in the current project, namely: (1.) Concerns that data might prove insufficiently robust given young Thai L1 student participants, and (2.) That the study calls for some way of identifying novel words familiar to test takers from word-internal cues. Like its VKS parent, the VSAT embodies self-assessment to facilitate rapid testing; it allows researchers to compile target (content) word lists of their choice; it incorporates verification procedures to address reliability concerns; and it claims sensitivity to the practicalities of conducting school-based research in busy educational settings. The VSAT ‘s lexical competence scale draws heavily upon that of the VKS, depicting a range of competence between zero knowledge (State 1) to an upper-end proficiency falling only marginally short of the native speaker[43] standard (State 6). State 2 does not properly represent a constituent of the scale as such but serves as a dummy category to which a test taker assigns target words s/he claims to know based upon prior knowledge (e.g. familiarity with affixes, or morphology). The full listing of VSAT knowledge states and brief descriptions is as follows:

State 1: I believe I have no knowledge of this word. (no verification)

State 2: I haven’t seen this word, but I think I know what it means. (no verification)

State 3: I have seen this word before, but I don’t know what it means. (no verification)

[43] Even if a word occupies State 6, this does not require that the test taker demonstrate native speaker understanding or usage skills.

State 4: I have seen this word before, and I think it means … (verification)

State 5: I know this word. It means … (verification: provide a synonym)

State 6: I can use this word in a sentence. (verification: provide a sentence)

State 1. (*I believe I have no knowledge of this word...)*

This state denotes zero knowledge –i.e. complete unfamiliarity with the target word. The test taker who allocates a word to this state claims no recollection of previous encounters with the word in written or oral form. The choice of "*believe*" (the VKS employs the term "*know*") acknowledges Durso and Shore's (1991) observation that those self-claiming no familiarity with meaning may nevertheless possess a measure of understanding.

State 2. (*I haven't seen this word, but I think I know what it means...)*

This represents a pseudo state (i.e., a dummy state); it does not denote a stage in word-meaning acquisition as such but filters out those words to which participants might assign meaning from drawing upon morphological, syntactic, or affix knowledge, i.e. information sources other than the contextual message.

State 3. (*I have seen this word before, but I don't know what it means...)*

This state holds words familiar to the standard of *noticing*, discussed in Section 2.3. Occupancy indicates that the word amounts intake that the learning system can both access and process. The test taker claims no familiarity with the word other than a possible previous encounter.

State 4. (*I have seen this word before, and I think it means* .....)

Word occupancy implies an accessible memory trace, or loose synonym, of the target vocabulary item. The test taker possesses only a limited receptive familiarity with meaning. Either (a.) s/he has a lesser understanding than would suffice to justify occupancy of State 5, or (b.) a comparable understanding to words in that state, albeit an understanding of which s/he remains unaware.

State 5. (*I know this word. It means ...)*

This state holds words sufficiently familiar that the test taker can employ them for the receptive purposes of reading and listening. The reader would understand the word if encountered during a conversation but cannot as yet supply it in syntactically and semantically well-formed clauses during writing or speaking.

State 6. (*I can use this word in a sentence...)*

This state captures aspects of productive knowledge, since it holds lexical items that the learner successfully incorporates into understandable (to the 'average,' literate, native speaker), semantically and grammatically correct utterances.

The VSAT does not supply a numerical measure of a test taker's knowledge of either a target word, or word corpus; it seeks only to identify the state to which a word belongs leaving the researcher to derive conclusions from the relative and absolute sums of target words 'occupying' the various descriptors. No division or summation of nominal data therefore occurs other than required to establish how many words lie within each of the 6 states. Like its VKS parent, the VSAT provides 'spot' measures of lexical competence at particular time points during a child's lexical development thereby allowing the researcher to apportion that competence into the knowledge categories the VSAT acknowledges. Following recommendations in Read (2000), the administrator allows verification question responses in either the L1 or L2.

3.4 VSAT administration[44]

VSAT testing begins with a preparatory stage during which the researcher compiles a list of target words (these s/he transcribes onto cards) along with a short familiarization session to acquaint the test taker with the test procedure. The choice of words (nouns, verbs, adjectives, or adverbs), their total, and the relative proportion of each word type, falls to researcher discretion the decision informed by study objectives and the practicalities of classroom-based research. A test session proceeds in the manner of the VKS with the researcher reading out a target word while handing the participant a card

[44] See Appendix 1 for administration details.

on which the word prominently appears. The card now in his/her possession, the test taker attempts to assign it to a state by placing the card on the relevant segment of a mat depicting each VSAT descriptor. Testing then continues according to the state the word currently occupies. If that state is 1, 2, or 3, the administrator presents the next target word from the list –the initial assignment presumed 'correct.' Placement in States 4, 5, or 6, however, triggers a 'two step' VSAT verification procedure. First, the administrator reads out a specified "corroborative" question (see Figure 3.3, below) to which the test taker obligatorily responds. Second, the administrator evaluates that response against the relevant success criteria (see pp. 76-78). Depending upon the outcome of that evaluation, the word either remains in the state to which the test taker originally allocated it or it undergoes reassignment to such alternative state that the administrator believes best captures the test taker's lexical understanding. The word placement now definitively established, the administrator moves on to present the next target word of interest. Testing continues in this same manner until every word has a VSAT state assignment.

	Gloss	Verification question
1	No knowledge of the word	N/A
2	Some familiarity with the word based upon morphology, affixes etc.	N/A
3	The word is familiar, but its meaning is not known	N/A
4	The student has some understanding of the word	Can you tell me *roughly* what this word means?
5	The word is described with semantic appropriateness; the student understands its meaning	Can you tell me *exactly* what this word means?
6	The word is used with semantic appropriateness and grammatical accuracy in a sentence	Can you put the word in a proper sentence?

Figure 3.3: VSAT verification questions and descriptive glosses of VSAT states.

For a word to remain in State 4, the participant must demonstrate a sufficient familiarity in response to the request "Can you tell me *roughly* what this word means?" An answer that suffices must include both details of a context in which the word appeared (e.g., the test taker claims to have encountered *canopic* from reading of ancient Egypt funerary arrangements) together with a broad indication of meaning that suggests, or hints at, an emerging understanding. Should the test taker satisfy this test, the word remains in State 4. A failure requires that the administrator allocate the word to State 3.

A word remains in State 5 only if the test taker orally supplies a precise synonym or definition either in his/her L1 or the target language, at the child's discretion (question: "Can you tell me exactly what this word means?"). The definition need not appear in a well-formed (i.e. syntactically error-free) clause but must demonstrate sufficient 'word familiarity' that the administrator entertains *no doubt* (i.e., is certain or sure) the test taker has comparable understanding to the average, literate, native speaker (see below). Should the synonym prove insufficient, the administrator allocates the word to State 4 or 3 –whichever of the two deemed most representative of the test taker's lexical competence.

A word retains a State 6 designation only if the test taker demonstrates productive use in a *grammatically* and *semantically* correct clause.[45] A clause displays grammatical correctness if the administrator considers any error (or errors) would not frustrate communication assuming native speaker language competence among the intended audience[46] –the test draws upon what Canale and Swain (1980) call communicative proficiency as opposed to technical accuracy of the sort typically defined in prescriptive school textbooks. A semantically accurate clause consists of a 'structure' that only attributes to the target word properties that correctly comprise part of its meaning specification; a child would fail this requirement, for example, if s/he described a 'pebble' as 'soft' or a 'table' as a device for cleaning floors. A further assurance that the word legitimately occupies State 6 derives from the VSAT stipulation that the test taker supply a target word synonym (either a word or clause) of comparable quality to that which the administrator might reasonably anticipate from a literate native

[45] Guidelines for this task appear in Appendix 1. The same requirement is incorporated into the VKS.
[46] The language capabilities of that native speaker, and how s/he might rate clauses for grammaticality, will vary from one administrator to the next. Should testing involve multiple administrators, a normalization session ensures all hold a common notion of that native-speaker standard.

speaker. This provision attempts to address the prefabricated language concern, noted above (p. 70). Should the administrator judge the synonym insufficiently native-like (see Figure 3.3) s/he obligatorily assigns the word to either State 4, or State 3 –the choice depending upon which objectively seems the more appropriate descriptor (see p. 76).

The following illustrative sentences help clarify the standards an administrator might apply in evaluating test taker responses. The target word is the italicized *colorless*:

1. The *colorless* thing look like glass you see through; then he drink it.
2. *The colorless was putted carefully on the top shelf.* (ungrammatical and non-semantic)
3. *The colorless liquid was a beautiful dark blue*. (grammatical but non-semantic)
4. *The colorless liquid mean you can't see when you look and this is so much.*

The first clause reasonably satisfies the semantic requirement (p. 76), correctly attributing to colorless the property of absence of color. The terms *like glass* and *see through* would likely strike the administrator as sufficiently conclusive. Grammatical correctness[47] appears plausible given a 'fairly' obvious intended meaning that comes across despite the 's' omission on the verb. Clauses 3 and 4 represent reasonable candidates for grammatical incorrectness, semantic incorrectness, or both. Sentence 2 would fail on both grammatical and semantic grounds given that colorless takes on the attribute of a concrete noun. Clause 3 fails the semantic test,[48] since here the test taker has assigned to colorless the property of 'blueness.' Clause 4, like Clause 3, fails the grammatical test given the native speaker's difficulty in establishing the meaning from among several possibilities. Should a test taker have supplied either Clauses 2, 3, the target word would most reasonably undergo assignments to State 3 with testing then proceeding as if initial allocation were to that state. Of the four clauses this leaves only Clause 1 as likely supplying sufficient evidence for the target word to remain in State 6, though

[47] "Likely" because this would depend upon the administrator's particular conception of native-speaker competence.
[48] Recall that the definition of native-speaker competence varies depending upon the understanding the administrator attaches to this term.

continued occupancy requires that the test taker supply a suitable synonym (see Appendix 1). Should s/he fail the synonym test the word would undergo reassignment to either State 4 or 3.

3.5 Three pilot studies of the VSAT

The VSAT, as a VKS variant, derives credibility from the same well-tested design features of its parent: e.g. it boasts a self-evaluation format capable of yielding robust data (Blachowicz & Fisher, 2006); it makes use of readily intelligible descriptors, employs a protocol for verifying assignment accuracy, and it embodies an easy-to-learn administration process. Even so, non-trivial concerns remain. Could, for example, the VSAT prove rather less suitable for use with the primary-aged children than with the tertiary students that its VKS parent attempts to assess? Can one reasonably assume that non-native English speakers possess sufficient skills to administer the VSAT in a manner that ensure it supplies robust and reliable data? Not least: How practical will VSAT administration prove in an authentic school setting?

The remaining sections of the chapter present the results of three pilot studies that seek to provide assurances that the VSAT will supply robust data given the intended deployment environment -a busy Bangkok based international primary school. The studies address the following questions:

1. How reliably will teacher assistants in the host institution administer the VSAT test instrument?
2. Does the VSAT yield stable ‘word-to-state assignments?’ That is, do essentially similar ‘word-to-state assignments’ arise if the test taker repeats the test assuming no relevant learning opportunities between the two testing occasions?
3. Do students reliably assign words to the self-assessed States 1 and 3?

3.6 Participants

Participants consisted of one or both of the following populations:

1. (Group 1) Primary-aged children currently studying in Year 4 classes at the host institution.
2. (Group 2) Teaching assistants employed as helpers to teachers in the Key Stage Two

primary year groups:

Group 1

The first group (n=10; four girls, six boys) consisted of a random selection of Thai L1 students from a population of 80 children attending the same private international school in Bangkok, Thailand. Participants' ages ranged from 8.8 to 9.7, with a mean of 9.2 (sd =2.4), comparable to the ages of children who would participate in the main study. Each student was receiving an education in an English only setting from English L1 monolingual instructors. Four children (three boys, one girl) had experienced all their primary schooling at the host institution while the remaining six represented transferees from other international schools in and around Bangkok. No child was currently receiving SEN support or had required such support at earlier points in their education; nor were any children disadvantaged by low socio-economic status.

Participants' L1 was Thai, with English exposure outside of school –based upon parental reports– limited to daily homework assignments and occasional supplementary instruction from private tutors. Each child at the time of participation was successfully achieving age-appropriate academic targets and had scored at least Level 3c in the SAT public examinations conducted during the 2007 academic year. None of the children admitted to having visited English-speaking countries other than for short excursions (two had visited the UK).

Group 2

The group comprised four teaching assistants who had expressed a willingness to conduct VSAT sessions during the main investigation. All were Thai L1 nationals though had benefited from many years of English instruction while enrolled in formal schooling. Two had gone on to major in English in local universities. Three held BA degrees and the fourth a BSc. Each assistant had served in their current roles for between three and five years at the host school, including at least one year working with Year 4 pupils. Self-reported scores on TOEFL tests ranged from between 590 and 600, sufficient for entry to most American university graduate programs. The qualifications and experience of participant volunteers compared favorably to those of colleagues not involved in the current study.

3.7 Experiment 1

Aim and rationale:

This experiment assesses inter-rater reliability. It aims to provide an assurance that potential VSAT administrators will evaluate responses to verification questions in a manner such that each will assign a target word to the same state as another. The experiment addresses the following two research hypotheses:

Issue 1

Research hypothesis: An assistant volunteer's interpretation of test-taker responses to VSAT verification questions (i.e. word to state assignments) is comparable to those of his/her peers.

$\boldsymbol{H_o}$: The sum of target words any one assistant assigned to a particular VSAT state is not *substantially* comparable to those of his/her peers (both unweighted and weighted[49] Kappa scores lie within the values 0–0.60).[50]

$\boldsymbol{H_a}$: The sum of target words any one assistant assigned to a VSAT state is substantially comparable to the word-to-state assignments of his/her peers (i.e. both unweighted and weighted Kappa scores lie within the values of 0.61- 1.00).

Issue 2

Research hypothesis: Assistant volunteers supply objectively correct word-to-state assignments based upon their evaluation of verification question responses.

$\boldsymbol{H_o}$: Assistants fail to assign a 'substantial' proportion (less than 85%) of target words to 'definitively' correct states.

$\boldsymbol{H_a}$: Assistants assign a 'substantial' proportion of target words (more than 85%) to definitively correct states.

[49] Weighted scores are also reported; the choice of unweighted scores for determining the status of H_o stems from the desire to limit type 1 errors –weighted scores were presumed likely to raise K values.
[50] Following Landis and Koch (1977), this corresponds to substantial to perfect agreement (see below).

3.7.1 Overview

The experiment involved a single child (henceforth, the test taker) who agreed to undertake a VSAT test session in the presence of all three assistants, the researcher acting as administrator. The assistants' received oral instructions explaining that they should independently evaluate the test taker's responses to verification questions following word assignment to States 4, 5, or 6 (see Section 3.5) and, following upon that evaluation, then independently allocate the word to the VSAT state s/he believed it correctly occupied. A Cohen's Kappa analysis of word-to-state assignments supplied the sought-after measure of agreement between one assistant's word placements and those of his/her peers (Question 1).

To address Issue 2, the study compares each assistant's determination of word placement with a presumed definitively correct assignment for each target word (see below). Again, a Cohen's Kappa analysis provided the measure of agreement rating.

3.7.2 Method

The experiment began with a normalization session to ensure assistants shared a common understanding of the verification question responses sufficient to justify a word retaining occupancy of the state the test taker had assigned it to initially. The session consisted of a discussion during which both assistants and researcher collectively evaluated the semantic and grammatical accuracy of a selection of children's responses gathered during the 2007 academic year while occupied with VSAT development. The session lasted approximately 45 minutes by which time the researcher and assistants had achieved the predetermined success criterion of four consecutive assignments upon which all agreed.

The test session employed a target word corpus (n=80) consisting of an equal number of nouns (n=20), verbs (n=20), adjectives (n=20), and adverbs (n=20). The words of each lexical class came from the following four frequency bands of word occurrence[51] (see Tables 4.1–4.4, below): (1.) five

[51] The source of word frequency was the BNC (Aston & Burnard, 1998).

words from the 1–1,999 band, (2.) five from the 2,000–2,999 band, (3.) five from the 3,000–3,999 band, and (4.) five from the 4,000–10,000 band. The selection process aimed to ensure at least some words would occupy each VSAT state, as they might during future VSAT sessions undertaken as part of the main investigation. Band 1 words stood as candidates for 'likely known,' Band 2 for 'less known,' and so on. The final target word corpus (see Tables 3.1–3.4) consisted of the following:

1. Group 1: Five nouns, verbs, adjectives, and adverbs randomly selected from the 1–999 most commonly occurring English words as identified in the BNC (Aston & Burnard, 1998).
2. Group 2: Five nouns, verbs, adjectives, and adverbs from the 1000–1,999 most common BNC words.
3. Group 3: Five nouns, verbs, adjectives, and adverbs from the 2,000–2,999 most common words in the BNC.
4. Group 4: A final five nouns, verbs, adjectives, and adverbs from the 4,000–10,000+ most common BNC words.

To select words for each group a random number generator supplied a figure within the numerical range of interest. If the number corresponded to a BNC content word, and the quota for that class remained unfilled, the word was added to the relevant corpus (a noun to the noun list, verb to verb list, and so forth).[52] In the event of 'no match' (e.g., if the number corresponded to, say, a preposition) or if the quota for the lexical class was full, the generator supplied alternative figures until such time that word placement in one or other class proved possible. Repeated cycles of number generation, referral to the BNC, and word allocation yielded the four 20-item lists of content words (of nouns, verbs, adjectives, adverbs) for use in the upcoming VSAT test session. The final words and their frequencies appear in Tables 1–4 below.

[52] The selection was not purely random, however. A word was ignored if, in the researcher's view, it would likely prove excessively time consuming for the child to explain or if (again in the researcher's view) the child might recognize it as a homonym for a word in another lexical class.

hand town team instrument year	writer guy illness symptom cloud	café outline rider horizon machine	spine headline grid marathon garlic
1–1,999	2000–2,999	3000–3,999	4,000+

Table 3.1: Nouns from the BNC, by frequency range.

great successful good large young	guilty angry rough dramatic eastern	curious weekly dull parallel random	cautious lazy grammatical abnormal glorious
1–1,999	2000–2,999	3000–3,999	4,000+

Table 3.2: Adjectives from the BNC, by frequency range.

suffer tell suggest begin encourage	assure tackle consult preserve sort	boost fetch conceal decorate forgive	explode smash spoil taste deposit
0–1,999	2000–2,999	3000–3,999	4,000+

Table 3.3: Verbs from the BNC, by frequency range.

sometimes clearly immediately already perhaps	gradually quietly gently initially occasionally	roughly typically lightly thoroughly nowhere	accurately independently repeatedly adequately beautifully
1–1,999	2000–2,999	3000–3,999	4,000+

Table 3.4: Adverbs from the BNC, by frequency range.

The test session involved administering the VSAT to a single student (the test taker) in the presence of three assistant volunteers. In the event of word allocation to State 6, 5, or 4, each assistant assessed the child's response to the relevant verification question, this read out by the assistant volunteers, taking it in turns to do so. Based upon that assessment the assistant either (1.) permitted the target word to remain in the original state, or (2.) allocated the target word to the state s/he considered more representative of the student's understanding. The task of recording each assistant's word-to-state assignments fell to the researcher.

Promptly following the test session, both 'assistant' participants and the researcher (henceforth, the '*panel*') reviewed each digitally recorded verification response to establish the definitively correct state of those words the test taker had assigned to States 6, 5, or 4. Any assistant who had allocated a target word to the correct state received a score of 1, while an incorrect assignment attracted a score of 0. The following formula supplied individual accuracy figures in percentage terms:

Accuracy = (total correct/total number of attempted assignments to States 4–6) x 100

To measure inter-rater reliability (Question 1), the study computes the Cohen's Kappa (*k*)[53] statistic (both the weighted and unweighted variants) for the word assignments of each possible pair of participants (i.e., Participant A's assignments with those of B and C, B with C and A, and C with A and B). Guidelines in Landis and Koch (1977) serve as a means to interpret Kappa scores (Figure 3.4, below).

Score Interpretation

0.0–0.20 Slight agreement

0.21–0.40 Fair agreement

0.41–0.60 Moderate agreement

0.61–0.80 Substantial agreement

0.81–1.00 Almost perfect agreement

Figure 3.4: Interpretive values for *k* (Landis & Koch, 1977).

[53] "*k*" values fall within the range of -1 to +1 and express the ratio of actual "matches" to what might arise from chance alone (the higher the "*k*," the greater the likelihood that raters truly concur).

To address Issue 2 (how definitively accurate were assistant assignments as opposed to whether they agreed with their peers), the study computes the percentage accuracy of each participant's word-to-state assignments employing the same accuracy formula given above. Again, a Cohen's Kappa analysis provides the indication of agreement between each participant and the *panel's* determination of correct assignments.

3.7.3 Results and discussion of Experiment 1

Experiment 1 sought to determine whether any single participant would assign words to the same state as their peers given the same verbally supplied test taker response to a VSAT verification question. Crosstab displays of participants' assignments and the results of the Cohen's Kappa analysis (unweighted and weighted) for the 35 words that the test takers attempted to place in States 6, 5, and 4 appear in Figure 3.5 below. Instances of agreement in word-to-state assignments for any 'participant pair' of interest present as figures within cells of the long diagonals (shaded). Those figures lying outside, i.e. above or below this diagonal, indicate cases of disagreement. For example, A and B both agreed that the same six words occupied State 3 albeit B assigned one target word to State 4 that C assigned to state 5.

Findings, Issue 1:

The unweighted Kappa analysis reveals k values ranging from a high of 0.65 between A and B (se: 0.099; 95% CI; 0.46–0.85) and A and C (se: 0.099; 95% CI; 0.46–0.85) to a low of 0.64 (se: 0.01; 95% CI; 0.45–0.85) between B and C (see Figure 3.5 below). These values all fall comfortably within Landis and Koch's (1977) band of substantial agreement. The number of disagreements between assistants amounted to nine, indicating an identical agreement of 76% of word-to-state assignments irrespective of the participant pairing. The average of the three unweighted Kappa statistics stands at 0.647, a figure again falling within the Landis and Koch (1977) substantial agreement range. A linearly weighted Kappa analysis[54] sensitive to the hierarchical order of the VSAT scale indicated higher agreement still, supplying *k* values lying between 0.80 (se: 0.063; CI:0.67–0.91) between Participants A and B, and 0.77 (se: 0.063; 95% CI: 0.63–0.91) for B and C. The values for all three participant pairings lie securely within the Landis and Koch (1977) score band signifying "almost perfect agreement" (Figure 3.4).

<u>*Determination*</u>: The null hypothesis is rejected (H_o: The sum of target words any one assistant assigned to a particular VSAT state was not 'substantially' identical to the state assignments of his/her peers).

Participant b (columns) × *Participant a* (rows)

		3	**4**	**5**	**6**
3		*6*	*1*		
4		*1*	*5*	*1*	
5			*2*	*5*	*3*
6				*1*	*10*

K=0.65(unweighted), SE=0.099
K=0.80 (weighted),SE=0.063

Participant c (columns) × *Participant a* (rows)

		3	**4**	**5**	**6**
3		*6*	*1*		
4			*6*	*1*	
5			*3*	*4*	*3*
6				*1*	*10*

K=0.65(unweighted), SE=0.099
K=0.79(weighted),SE=0.063

Participant c (columns) × *Participant b* (rows)

		3	**4**	**5**	**6**
3		*6*	*1*		
4			*7*	*1*	
5			*1*	*3*	*3*
6			*1*	*2*	*10*

K=0.64(unweighted), SE=0.10
K=0.77(weighted),SE=0.07

Figure 3.5: Cross comparisons of participants' word-to-state allocations of target words.

<u>Findings, Issue 2</u>:

Table 3.5 depicts the percentage accuracy scores for each of the participant volunteers to

[54] These take into account that the states lie within a hierarchical "order" running from 3 to 6.

provide a simple visual measure of agreement between word-to-state assignments and the (presumed) definitively correct scores as determined by the aforementioned *panel* (p. 84). Though accuracy proved variable, sums of correct scores appear, arguably, impressive. The lowest accuracy figure amounted to 82.9%, recorded for Assistant C, and the highest, 88.6% for Participants A and B alike. The mean assignment accuracy for the 'participant group' amounted to an encouraging 87% (sd=.57), corresponding to an average of just 4.6 misplacements of the 35 words the test taker had assigned to States 4-6 initially. The researcher's score of 95% equates to two words misplaced, the errors in each instance stemming from incorrectly interpreting verification question responses the test taker had supplied in the L1 –a reminder of the need for truly bilingual administrators. In just two cases did more than one participant assign the same word to the same incorrect state: A and C assigned 'rough' (adj.) to State 5 when it properly occupied State 6, while B and C assigned 'gently' (adv.) to State 4 while the *panel* placed it in State 6. Assignment errors proved typically unique, each assistant misplacing different words from those of their peers. Lexical class and 'number of errors' seemed similarly unrealted. Of the 14 words the group misplaced,[55] three were verbs (suggest, sort, begin), three nouns (instrument, writer, cloud), three adjectives (rough x 2, and angry), and four adverbs (immediately, gently x 2, beautifully, perhaps). The most words of the same lexical class any one participant misplaced consisted of the three adverbs incorrectly assigned by Participant C. Individual accuracy scores, and agreement ratings, become more impressive still if test takers indeed correctly assigned target words to States 1–3 (the likelihood of this receives attention below). For Participant A, assignment accuracy increases from 88.6% to 95%; for B, assignment accuracy rises to 95%; and for C, to 92.5%.

The Cohen's Kappa agreement ratings between the *panel* (above) and participants appear in Figure 3.6. While the figures indicate some variation, they prove consistently high. The agreement between the *panel* and Participant B stands out as striking (k=0.84; se: 0.073; 95 % CI: 0.699–0.98)

[55] This refers to the sum of words misplaced by all three participants.

though only marginally more so than that observed between the *panel* and A (*k*=0.81; se: 0.080; 95% CI: 0.64–0.96). The least agreement lay between the *panel* and C though this still amounts to an impressive conformity at *k*=0.76 (se: 0.09; 95 % CI: 0.59–0.93). The *k* values all fall comfortably within the Landis and Koch band of "almost perfect agreement" (as we see in *panel* versus B, and *panel* versus A) and "substantial agreement" (*panel* versus C). The linear weighted Kappa values emerge as more impressive still. The highest weighted *k* value stands at 0.90 (se: 0.04; 95% CI: 0.82–0.99), this for the *panel* and B, while the lowest is 0.86 (se: 0.05; 95% CI: 0.75–0.96) between the *panel* and C. These figures lie within the 'almost perfect' agreement band of the Landis and Koch interpretive criteria. The mean weighted value is *k*=0.88.

Determination: The null hypothesis is rejected (H_o: Assistants failed to assign a 'substantial' proportion (less than 85%) of target words to 'definitively' correct states).

Participant	% of words correctly assigned	Number of words incorrectly assigned	Words incorrectly Assigned
A	31/35 = 88.6	4	immediately (adv), rough (adj), suggest (v), instrument (n)
B	31/35 = 88.6	4	writer (n), sort (v), begin (v), gently (adv)
C	29/35 = 82.9	6	angry (adj), beautifully (adv), gently (adv), perhaps (adv), cloud (n), rough (adj)

Table 3.5: Participant 'word-to-state' accuracy scores.

	Participant A				
		3	**4**	**5**	**6**
Panel	**3**	*6*			
	4	*1*	*6*	*1*	
	5		*1*	*7*	
	6			*2*	*11*
	k=0.81 (unweighted), se=0.080; k=0.88 (linear weighted), se=0.05				

Participant B					
		3	**4**	**5**	**6**
Panel	**3**	*6*			
	4	*1*	*7*		
	5		*1*	*6*	*1*
	6			*1*	*12*
	k=0.84 (unweighted), se=0.07;k=0.90 (linear weighted), se=0.04				

	Participant C				
		3	**4**	**5**	**6**
Panel	**3**	*6*			
	4		*8*		
	5		*2*	*4*	*2*
	6			*2*	*11*
	k=0.76 (unweighted), se=0.09 K=0.86 (linear weighted), se=0.05				

Figure 3.6: Cross comparisons of participant and *panel* word-to-state allocations.

3.7.4 Conclusion of Experiment 1

The experiment demonstrates both substantial inter-rater agreement (Issue 1) and impressive accuracy in participant word-to-state assignments given the panel's (presumed) definitively correct assignments (Issue 2). The findings provide a powerful affirmation that assistants willing to participate in the main investigation (Chapter 4) will administer the VSAT in a manner yielding objectively accurate data. That children took little time assigning words to states (rarely did they require more than 5 seconds or so) allayed concerns that testing might prove overly time consuming for classroom-based research. Moreover, assistants' promptness in evaluating verification question responses –they typically completed the task in a matter of seconds– implied feasible VSAT administration given the limited time available within the school's standard 35-minute lesson format. All assistants expressed confidence in their ability to interpret test-taker responses to verification questions and assign words to the appropriate state. By general agreement testing proved undemanding for administrators and test takers alike.

The highest scorer from the Experiment 1 (Question 2) investigation (Participant B) was invited to conduct VSAT administration sessions in the main study. Both A and C expressed willingness to assist as may be required, such assistance to involve helping the researcher evaluate verification responses children elected to deliver in Thai, along with supervising RR sessions.

3.8 Experiment 2

Aim and rationale:

This experiment involved four (student) participants in a test/retest investigation that assessed the stability of children's word-to-state assignments over two successive testing occasions. The between-test interval amounted to seven days,[56] a period deemed sufficiently short to rule out learning between test administrations yet long enough to discount children recalling word-to-state assignments from Test Occasion 1. The experiment tests for reliability –the capacity of an instrument to supply the same data of word-to-state assignment over successive occasions (Carmines & Zeller, 1979). The investigation addresses two Research Issues:

Issue 1

Research hypothesis: Test takers, as a group, display stable assignments at each test occasion (word-to-state placements at Test occasion 1 corresponding to placements at test occasion 2).

H_o: Test takers, as group, display *insufficient* agreement in word-to-state assignments (less than 90% correspondence in word-to-state allocations).

H_a: Test takers, as a group, display *substantial* agreement in word-to-state assignments (90% or more agreement in word-to-state allocations).

Issue 2

Research hypothesis: The stability scores of any single participant are not markedly more, or less, variable than those of any other participant.

[56] These compare with Waring's (2000) 3-day interval and Paribakht and Wesche's (1993) 14.

H$_0$: For any single test taker, word-to-state assignments at Occasions 1 and 2 prove essentially dissimilar (Kappa scores of 0–0.60).

H$_a$: For any single test taker, word-to-state assignments prove highly similar (Kappa scores of 0.61–1).

3.8.1 Method

The experiment employs a '32 target word corpus' drawn from each of the four BNC-based frequency bands employed in Experiment 1. Again, the selection sought to replicate as far as practicable the testing conditions (exposure to target vocabulary ranging from the common to relatively rare) that children would encounter during the main study (see Chapter 4). The words from each band consisted of an equal number (n=8) of nouns, verbs, adjectives, and adverbs. The same random number-generator procedure employed in Experiment 1 supplied the words of interest. The final corpus, with words apportioned by lexical class and frequency band, appears in Table 3.6 below.

able, whole (adj) suffer, tell (v) method, claim (n) right, quickly (adv)	*grateful, favorite (adj) defeat, capture (v) mouse, lane (n) secondly, primarily (adv)*	*full-time, unhappy (adj) pop, beg (v) chapel, oxygen (n) forever, backwards (adv)*	*cruel, profitable (adj) grip, pronounce (v) statue, sauce (n) rightly, honestly (adv)*
1–999	*2000–2999*	*3000–3999*	*4,000+*

Table 3.6: A random selection of BNC content words, drawn from four frequency bands.

A day prior to the experiment children attended a familiarization session to explain the VSAT test procedure and provide an opportunity to address participant concerns. The occasion doubled as a chance to inform potential participants that testing had no bearing upon schoolwork and remind them that they could exercise the option of withdrawing at any timepoint. The first tests took place during the following two days and repeat tests seven days thereafter, the researcher acting as test administrator on each occasion. To limit the confounding effects from external factors (lighting, tiredness, etc.), tests occurred under as identical conditions as reasonably possible –e.g. at the same time during the school day (during the lunch break), in the same location (the researcher's classroom), and in the presence of the same supervisory staff members. Each test session took approximately 12 minutes to complete.

For analyzing and interpreting data, the study uses crosstab displays to supply simple visual impressions of word-to-state stability between testing occasions. The configuration of Test 1 and Test 2 scores along the columns and rows in such displays helpfully allows for three types of deduction (see Figure 3.7):

1. Figures that appear within the diagonal marked by Cells 1, 7, 13, 19, and 25 depict the total number of words that remained in the same state during both testing occasions.
2. Figures to the right of the diagonal (indicated by '+' signs) signify words that relocated to higher states than occupied at test Occasion 1.
3. Numbers to the left of the diagonal (depicted with '-' signs) indicate word movements to numerically lower states.

O[1]	+[2]	+[3]	+[4]	+[5]
–[6]	O[7]	+[8]	+[9]	+[10]
–[11]	–[12]	O[13]	+[14]	+[15]
–[16]	–[17]	–[18]	O[19]	+[20]
–[21]	–[22]	–[23]	–[24]	O[25]

Figure 3.7: A crosstabs display of Test 1 and 2 performance.

To measure the stability of participant assignments (the Issue 1 objective) the experiment employs the same Cohen's Kappa test described in reference to Experiment 1, the data for the present test/retests derived from the particular word-to-state placements children supplied during the first and second test sessions. Again, the Landis and Koch (1977) criteria provide the basis for interpreting the *k* statistic (see Figure 3.4).

To address Issue 2 required comparing each participant's Test 1 and Test 2 word assignments to determine the extent of relocations from one state to another. This involves comparing-participant A's assignments at Test Occasion 1, for example, with his/her assignments at Test Occasion 2, repeating this with Participant B's assignments at each test occasion, and so on for the remaining participants. Crosstab displays reveal (a.) the magnitude and (b.) the direction of word movement (whether from higher to lower states) between test occasions.

3.8.2 Results and discussion of Experiment 2

<u>Findings, Issue 1</u>:

The experiment asked whether successive VSAT test sessions with the same test taker yield essentially equivalent word-to-state assignments in the absence of opportunities for learning during the between-test interval. The data from the test/retest scores appear in Figure 3.8 (below); the figures in brackets express totals of unchanged words in percentage form (i.e., as a proportion of 128).

Of the total words, 111 (or 86.7%) remained in the state they occupied at Test Occasion 1, the sums appearing in the diagonal running from Cell 1 to Cell 15. The highest proportion of 'stable' words by state was 96% (the 22 words that retained State 1 status) and the lowest, 80% (the four words occupying State 2). Of the 17 words (or 13.2%) that relocated, 12 (or 70%) moved to a state adjoining the one occupied at the first test occasion. Thirteen transferred to a numerically lower state and four to a higher. The unweighted Kappa statistic amounts to an impressive 0.862 (se: 0.035; 95% CI: 0.793–0.930), a figure that falls comfortably within the upper end of Landis and Koch's (1977) range for substantial agreement (Figure 3.5).

<u>*Determination*</u>: The null hypothesis presumption of 'insufficient' agreement is rejected (H_o: < 90% correspondence in word-to-state allocations).

Time 2							
		S1	S2	S3	S4	S5	S6
Time 1	S1	22(96%)	1				
	S2	1	4(80%)				
	S3		2	24(83%)	2	1	
	S4			1	21(95%)		
	S5			2	2	25(93%)	
	S6			2		3	15(83%)

Figure 3.8: Number of words placed in each state at Test Occasions 1 and 2.

Findings, Issue 2:

The crosstab displays for individual participant's word assignments (Test 1, Test 2 assignments) appear in Figure 3.9 below. The unweighted *k* values lie between a high of 0.960 (se = 0.038; 95% CI: 0.885–1) for Participant A and a low of 0.760 (se = 0.08; 95% CI: 0.590–0.933) for Participant D. For participants, B and C, the Kappa values amounted to 0.805 (se: 0.080; 95% CI: 0.648–0.962) and 0.804 (se = 0.080; 95% CI: 0.690–0.985), respectively. In all four test/retest cases, *k* lies within the Landis and Koch (1977) range of substantial to near perfect agreement. The weighted Kappa values proved rather more impressive still, falling between a high of 0.961 (A) and a low of 0.857 (C) signifying almost perfect test/retest agreement[57] (Landis and Koch, 1977). The likelihood of words shifting from the state occupied at test occasion 1 proved broadly independent of initial state assignment (see Figure 3.9). For Participant A the most volatile state emerged as State 3, which lost a word to State 5 at test occasion 2. For Participant B, State 6 proved the most volatile while for Participants C and D, most word relocations were from States 5 and 6, respectively. The findings reveal

[57] On a cautionary note, one might question whether State 2 really falls into the measurement scale as such. This should be borne in mind when interpreting the weighted statistic.

no obvious association between the sum of word-to-state relocations and the particular test 1 word-to-state placements.

Nor do the data imply a relationship between likelihood of movement and part of speech. For Participant A, a sole noun relocated during the seven days between testing occasions. For Participant B, a noun relocated along with two verbs, one adjective, and one adverb. For Participant C movement involved two nouns, two adverbs, and one adjective, while for Participant D, two adverbs relocated along with two adjectives, one verb and one noun. With so few shifts between test occasions, the percentage stability scores for each participant (the percentage of words that remained in the original state) proved impressive: 97% for A, 84% for B, 87% for C, and 81% for D.

<u>*Determination*</u>: The null hypothesis is rejected (H_o: For any single test taker, word-to-state assignments at Occasions 1 and 2 were essentially dissimilar (Kappa scores of 0–0.60).

	Participant A, test 2					
Participant A, test 1	1	2	3	4	5	6
1	10					
2		1				
3			5		1	
4				6		
5					4	
6						5

	Participant B, test 2					
Participant B, test 1	1	2	3	4	5	6
1	3					
2	1					
3		1	6			
4				5		
5				1	6	
6			1		1	7

	Participant C, test 2					
Participant C, test 1	1	2	3	4	5	6
1	6					
2		1				
3		1	6	1		
4				6		
5			2		6	
6			1			2

	Participant D, test 2					
Participant D, test 1	1	2	3	4	5	6
1	3	1				
2		2				
3			7	1		
4			1	4		
5				1	9	
6					2	1

Figure 3.9: Word-to-state assignments for each participant.

Participant	Word class	Word	Relocated from…
A	Noun	progress	State 3 to State 5
B	Noun Verb Verb Adjective Adverb	chapel pop defeat full-time rightly	State 6 to State 3 State 3 to State 2 State 5 to State 4 State 2 to State 1 State 6 to State 5
C	Noun Noun Adverb Adjective Verb	claim oxygen primarily profitable suffer	State 3 to State 2 State 3 to State 4 State 5 to State 3 State 5 to State 3 State 6 to State 3
D	Adverb Adverb Verb Noun Adjective Adjective	forever honestly beg chapel full-time able	State 1 to State 2 State 3 to State 4 State 4 to State 3 State 5 to State 4 State 6 to State 5 State 6 to State 5

Table 3.7: Word-to-state relocations between each test occasion by participant.

3.8.3 Conclusion of Experiment 2

The above findings indicate stable, systematic, results over a seven-day period. The figures for participants both as a group (Issue 1) and at the individual level (Issue 2) compare favorably with the test/retest consistency scores Paribakht and Wesche (1996) reported for the VKS from a two-week interval between test sessions, and those Waring (2000) recorded for the State Rating Task.

3.9 Experiment 3

Aim and rationale:

To determine how reliably participants assigned words to States 1 and 3, neither state requiring administrator verification. The experiment addresses two Issues:

<u>Issue 1:</u>

Research hypothesis: Participants correctly assign ‘unknown’ words to State 1

H_o: Participants, whether as an indivisible group or individually, fail to allocate more than 85% of genuinely ‘unknown’ words to State 1.

H_a: Participants, whether as an indivisible group or individually, allocate substantially all unknown (>85%) words to State 1.

Issue 2

Research hypothesis: Participants correctly assign words to State 3

H_o: One or more participants failed to allocate at least 75% of words that should properly occupy State 3 to that state.

H_a: Each participant correctly allocated more than 75% of words that should properly occupy State 3 to that state.

3.9.1 Method

Ten child participants each undertook a single VSAT test session during Term 1 of the 2009 academic year. Target words (n=100) consisted of three types. Type 1 (see Table 3.8) comprised 30 non-words conforming to English phonotactic and graphophonic rules but auditorily dissimilar to authentic Thai or English vocabulary children might reasonably have familiarity with. These researcher designed words stood as obvious choices for test taker assignment to VSAT State 1. Type 2 words (n=30) comprised an equal number of nouns, verbs, adjectives, and adverbs, that would most plausibly occupy State 3. Identifying these words involved two steps: (1.) Compiling a preliminary list of candidate items –words that participants might reasonably have encountered but sufficiently infrequently and under circumstances that they likely remained unavailable for language use (it fell to the researcher, drawing upon familiarity with Year 4 Thai L1 students, to compile this list); and (2.), testing each word for suitability as a State 3 term. A total of one hundred and fifty such words proceeded to the testing stage of the selection process.

The testing (step 2) sought to identify those words that children had likely *noticed* during reading or other language exposures. This involved inviting three Year 4 children to rate all 150 vocabulary items for familiarity by independently matching each to a list of descriptive captions borrowed from Zimmermann (1997); namely: (a.) “I don’t know the word,” (b.) “I have seen the word before, but I am not sure of the meaning,” (c.) “I understand the word, but I do not use it,” and (d.) “I can use the word in a sentence.” Those words that received a caption ‘b’ rating (total = 38) entered a pool of candidate State 3 (i.e., possible ‘*noticed*’) items. A random selection of 30 of these items supplied the particular Type 2 words employed in the experiment (see Table 3.9).

Non-words (to be assigned to State 1)		
mork, toker, bettle, *cadly, nased, yoot,* *mandly, casle ,blund, mear*	*greal, bick, prink, trilp* *gragly, palk,* *tance, vack, rimple,* *parrow*	*molden, tring, tantic,* *bettle, nase, mand,* *gadged, flink sind,* *jurg*

Table 3.8: Non-words, Experiment 3.

Candidate State 2 words		
basilica (n) *canopic (adj)* *water-butt (n)* *appropriately (adv)* *vizier (n)* *translucent (adj)* *pedagogue (n)* *therapy (n)* *trachea (n)* *resume (v)*	*centurion (n)* *iceni (n)* *isis (n)* *transpire (v)* *watershed (n)* *phloem (n)* *filament(n)* *abrasive (adj)* *streamlined (adj)* *chord n)* *duet (n)* *anemometer (n)*	*conscientiously (adv)* *recollect (v)* *abridged (adj)* *tonic (n)* *republic (n)* *compile (v)* *associate (v)* *assiduously (adv)*

Table 3.9: Candidate ‘*noticed*’ words.

A further word type, Type 3 words (see Table 3.10), consisted of 40 distracters bringing the total number of vocabulary items for testing to 100. These additional words acted as ‘fillers’ and functioned to add realism to the test-taking experience by raising the likelihood that all VSAT states would receive some word entries. The particular Type 3 words comprised a random selection from the 80-item list employed in Experiment 1 (Tables 3.1–3.4).

Type 3 words
adequately, assure, café, cloud, conceal, consult, deposit, dull, eastern, garlic, good, grammatical, great, guilty, illness, instrument, large, lazy, lung, machine, most, occasionally, perhaps, random, repeatedly, rider, signal, smash, sort, spoil, suffer, suggest, symptom, taste, tell, thoroughly, treaty, typically, writer, young.

Table 3.10: Random words from the BNC.

To address Issue 1 (How accurately did children assign words to State 1?), the study identifies the number of Type 1 words (presumed unknown) participants allocated to that state as a proportion of the Type 1 total. The percentage of correct State 1 assignments is:

$$\left\{ \frac{\text{total words of Type 1 assigned to State 1}}{\text{30 (the total non-words employed in the test)}} \right\} \times 100$$

A binomial sign test (one-tailed; $p<0.05$) of each participant's scores identified the likelihood of any participant's sum of correct assignments having arisen from purely random word-to-state placements.

Issue 2 (p. 97) required establishing those words participants correctly assigned to State 3 (i.e., words s/he could claim to have *noticed*) from the total s/he attempted to allocate to that state. To identify a correct State 3 word placement, the study employs two tests of familiarity:

Test 1:

Could the participant supply a likely context where s/he might have previously encountered the target word? A test taker would satisfy this requirement, allowing a presumption of *noticing*, if, for example, s/he claimed to have heard *centurion* during a discussion on ancient Rome, or the word *duet* during a music lesson.

Test 2:

Could the participant describe an association between the target word and some practice, entity, belief etc. with which it carries an association. A participant satisfied this test if s/he associated, say, *lung* with the act of breathing, or *conceal* with hiding.

Should the test taker fail either test, or the researcher still entertain reasonable 'evidence based' doubts concerning State 3 occupancy then the word underwent obligatorily assignment to State 1, the only reasonable alternative. While this assignment potentially inflates the sum of State 1 words (i.e. some genuine State 3 words will receive State 1 designations), and lowers the likelihood of null-hypothesis rejection, it offers a welcome assurance of sorts that words identified as State 3 occupants truly represent vocabulary familiar to the *noticing* standard. Since vocabulary other than the candidate *noticed* words (Table 3.9 above) might occupy VSAT State 3 (the possibility remained that children possessed some familiarity with certain Type 2 words), the proportion of correct word assignments to State 3 derives as follows:

$$\left\{ \frac{\text{Total words that retained their State 3 status (given Tests 1 and 2 above)}}{\text{Total words of Type 2 and Type 3 that the test taker attempted to place in State 3}} \right\} \text{x } 100$$

The investigation employs a binomial sign test (one-tailed; $p<0.05$) to establish the probability that the proportion of correct assignments for any test taker conceivably arises from pure random chance.

3.9.2 Results and discussion, Experiment 3

<u>Findings, Issue 1</u>:

The experiment asks how accurately young (9 year old) children assign words to those VSAT states that do not elicit verification questions. The findings appear in Table 3.11, which displays both the number and percentage of the 30 Type 1 words each participant correctly placed in State 1 during VSAT testing.

The table reveals low variability in children's assignment behavior, with the highest scoring among the child volunteers (Participants 5, 8 and 10) each achieving 29/30, a 97% accuracy (Table 3.8). The least successful assigners emerged as Participants 2 and 6, each of whom achieved 87% accuracy corresponding to raw scores of 26/30. The median score for all 10 children came to a substantial 93% and the mean, 92.4% (sd.3.86). A binomial sign test (one-tailed; sig. $p<0.05$) to establish the likelihood of a child having achieved his/her score by 'chance' indicated a probability of zero, the test returning identical values of $p=0.00$ for each participant. The collective, or overall, accuracy of the 10 participants (or 'indivisible group') proved no less impressive. Of the total 300 potential word placements in State 1 (i.e. 10 x 30), the children correctly assigned 277 words to that state, or just 23 to a state other than State 1.

The causes of misassignment emerged in a post-test discussion during which each child explained their State 3 placement for those words allocated to that state but which the researcher felt more properly belonged in State 1. This feedback indicated a common factor: a sense on the child's part of having encountered the target word on some prior occasion. The following quotes come from the discussion session:

"*I think I read it somewhere." (Participant 3)*

"I'm sure it is a word!" (Participant 5)

"I think I hear people say [it] sometimes, but I don't know what it mean[s]." (Participant 6)

Invited to elaborate, children attributed this familiarity to either perceived similarities in the orthography between target non-words and words they regarded as known (one child claimed the string *tring* had appeared in the English version of a Thai comic), or auditory similarities –for example, the non-word *mork* and the Thai verb corresponding with the English '*to tell*.' The possibility that successive encounters with the same unknown word(s) during VSAT testing might itself have the adverse effect of inducing a sense of familiarity prompted a revision to the VSAT administration procedure for longitudinal studies that call for periodic testing with the same target word corpus. The revision, subsequently adopted, required that a word assigned to State 1 during a VSAT session would obligatorily retain occupancy of that state irrespective of the state in which a test taker might place it during subsequent test occasions.

Participant	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Percentage	90	87	90	93	97	87	93	97	93	97
Raw score	27	26	27	28	29	26	28	29	28	29
Significance (p-value)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3.11: The proportion of State 1 words correctly assigned by 10 participants.

<u>Determination</u>: The null hypothesis is rejected (H_o: Participants, whether treated as an indivisible group or individually, fail to allocate more than 85% of genuinely ‘unknown’ words to State 1).

<u>Findings, Issue 2</u>:

Table 3.12 (below) displays each participant’s percentage accuracy scores for target words of Type 2 or 3 allocated to State 3 following upon the child having satisfied the relevant verification tests (p.99). The total words a child attempted to assign to this state varied given individual differences in the number of *noticed* Type 2 words, and that Type 3 words proved familiar to varying degrees depending upon a participant’s previous language exposure. The total words a child sought to place in State 3 ranged from 18 (Participant 1) to a high of 28 (Participant 7).

Assignment accuracy ranged from 95% (20/21) from Participant 10 to a low of 83%

(Participant 3), corresponding to a raw score of 19/23. The probability of a child achieving his/her score from chance proved low with binomial sign tests (one tailed; $p<0.05$) placing this likelihood at between 0.000% for Participant 9 to a 'high' of 0.001% for Participant 1. For the participant group as a whole, the average assignment accuracy amounted to 88.7%, the probability of this arising from random word placement (i.e. that children correctly assigned a word to State 3 having satisfied the tests) emerging as a lowly 0.001%. The sum of words inadvertently misassigned to State 1 during the test sessions given the Type 1 bias (i.e., the likelihood of the researcher placing some words in State 1 despite that the child had indeed *noticed* them) remains unclear. Time constraints made it impractical to conduct post-tests to identify any possible misplacements.

Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10
17/18	23/26	19/23	18/21	20/24	17/18	24/28	23/26	19/21	20/21
94%	88%	83%	86%	83%	94%	86%	88%	90%	95%

Table 3.12: The proportion words correctly assigned to State 3 by 10 participants.

Determination: The null hypothesis is rejected (H_o: One or more participants failed to allocate at least 75% of words that properly belong to State 3 to that state).

3.9.3 Conclusion of Experiment 3

The two investigations establish that children accurately assign target words to either State 1 or State 3 from unaided introspecting upon their lexical understandings, a finding as valid for individual participants as for the participant group. The results argue against amending what would appear to be a serviceable and reliable self-assessment test instrument and imply that more extensive verification procedures would insufficiently add to VSAT robustness to justify the longer test sessions and/or interruptions to school routines that this necessarily entails. That some misassignments will emerge during testing remains a strong possibility. It would seem, however, that these will not compromise VSAT reliability.

3.10 Interviews with assistant participants

A post-test discussion attended by the researcher and the assistant volunteers took place one week after the conclusion of Experiment 3 to evaluate VSAT practicality in classroom settings. Results from this feedback proved reassuring with assistants variously describing the VSAT as “straightforward,” “simple,” and “intuitive” to administer. All agreed that children felt comfortable with the testing process and willing to answer verification questions. A sense that test administration time might prove excessive, however, given the anticipated time available during the main investigation prompted the suggestion that children complete a paper-based assessment rather than undergo teacher/student VSAT sessions. The proposal, agreed by all, called for a child to draw lines connecting each target word to the appropriate VSAT descriptor. An administrator would then proceed to verify those assignments as soon as practicable thereafter. A small-scale trial of this option with Year 4 children proved successful prompting the decision to employ such a ‘paper’ test in the main study.

3.11 Concluding points

Experiment 1 provided the sought-after statistical reassurance bilingual school staff from the host institution could indeed administer the VSAT in a manner yielding comparable results to a *panel* comprised of themselves and the researcher. Each assistant participant therefore duly received an invitation to assist in conducting VSAT administration sessions for the proposed 2009 and 2010 investigations (Chapter 4 states the particular capacities in which each contributed). The experiment affirmed the need for ‘strong’ English language skills among candidate administrators but also highlighted the need for native-like familiarity with Thai in order to interpret verification question responses supplied in the child’s L1. Perhaps surprisingly, given the student body, participants rarely chose the L1 option despite awareness they could do so. Why this was so, remains unclear –Possibly children felt reluctant to resort to Thai given the school's English promotion policies. Notably, the duration of testing, per target word, amounted to only marginally less when children reverted to L1 use.

Experiment 2 indicated relatively stable student word-to-state assignments over the four days between test sessions, suggesting a pattern of principled and consistent test-taker responses. The test/retest data compares favorably with that for the VKS and provides strong affirmation that the VSAT should supply sufficiently reliable data given the aims of the current investigation. Indirectly, the test provided a 'guarantee' of sorts that the VSAT neither induces stress among child participants nor raises anxiety among those adults charged with administration.

Findings from Experiment 3 dispel concerns children might inaccurately assign words to states that did not prompt the verification procedure (namely, States 1 and 3). All child participants made reliable word placements with assignment accuracy measures generally proving comparable whether test takers chose to place words in State 1 or 3. Children (interestingly) reacted to non-words in the same manner as when encountering authentic, yet unfamiliar, English vocabulary (no queries directed to the researcher, no undue puzzling over the word card but simply a focus on correct placement). The time a child expended in word placement to State 1 rarely exceeded a few seconds offering an assurance the VSAT administration might not prove overly time consuming to administer. Placement in State 3 took marginally longer, though not sufficiently so as to raise concerns regarding VSAT classroom use.

3.12 Summary

This chapter began by evaluating the VKS as a candidate for use in the current research project. That discussion revealed several difficulties that arise when researchers attempt to reconcile the often competing demands for robust data on the one hand, calling for more demanding verification procedures, and the practicalities of conducting an investigation in an authentic school setting[58] on the other. While the VKS appears unsuitable for use given the aims of the current study, the chapter argued that an adapted version, the VSAT, could serve as a viable and robust alternative. The results from three pilot studies proved reassuring and allow for a confident assertion the VSAT is well-suited for exploring the research issue and research questions presented in Section 1.3 (see also, Section 4.2.1 below). The following chapter describes the VSAT-based research methodology of the dissertation's main investigation.

[58] As noted earlier, that the host school had an 'English language only' policy may have discouraged children from answering VSAT verification questions in their L1.

Chapter 4

Methodology 2 (The study design)

4.1 Introduction

The previous chapter described efforts to develop a data-elicitation instrument (the VSAT) for use in classroom-based settings with child participants then went on to report upon the design of three small-scale investigations the results from which provided assurances that the VSAT supplies reliable data from which to explore issues of concern to the present study. The current chapter begins with a review and clarification of the Research Questions presented in Chapter 1. The discussion then moves on to describe a detailed and comprehensive VSAT based methodology to address each question in turn.

4.2 A brief methodological overview of the study

This investigation traces participants' (n=28) acquisition and retention of novel vocabulary items encountered during reading sessions that replicate as far as practicable children's experiences during authentic school-based RR study periods. The dissertation addresses a single issue: How do the time intervals (measured in days)[59] between reencounters with the same novel word during classroom RR impact upon learning outcomes in terms of reader uptake of new word meanings? To explore this question the study requires that each participant read five sets of texts during timetabled RR lessons,[60] the sets so designed to incorporate four unique (to that set) target non-words, each of which appears in either its base or inflected form 12 times in that same set. A reading schedule (see Section 4.19 below) specifies when reading occurs and ensures that a set exposes the participant to its non-words under a

[40] Days, or rather daily RR sessions (of 35 minutes duration), serve as the unit of measurement, because (a) teachers tend to organize themselves around these time units and (b) other time periods were impractical given the timetabling of the school day.

[60] Each child read the same set of texts.

uniquely massed or distributed learning condition. A child reading Set 1, for example, encounters the 12 repetitions of each of the four target words[61] during a single (35 minute) reading session; Set 2 exposes the reader to those repetitions (12) over two consecutive daily (35 minute) sessions, Set 3 over three sessions, and so on for the remaining sets. Having completed a set of texts, a child promptly undertakes a VSAT test of the four embedded non-words. The current study records, analyses and discusses the sums of target words children learned from reading each set, drawing upon VSAT elicited indications of target word uptake. Because sets of texts display a fundamental similarity in regard to genre, syntactic complexity, writing style and lexical attributes, differences in word learning from reading experiences, reasonably stem from the particular presentation time over which target vocabulary encounters took place. The frequency of encounters with the target words of each set, the duration over which reading occurs, and the number of books appear in Table 4.1 below.

Sets of texts	*Number of books in the set*	*Reading duration*	*Number of times each target word appears in each book*
Set 1	1	1 day	12
Set 2	2	2 days	6
Set 3	3	3 days	4
Set 4	4	4 days	3
Set 5	5	5 days	2.4

Table 4.1: Gross features of experimental sets of texts.

4.2.1 The Research Questions

Chapter 1 defined spaced learning, and suggested this represents a construct teachers might potentially exploit to maximize vocabulary learning during RR sessions. That children learn most

[61] That is, the child will encounter each of the four distinct target words embedded in that set 12 times.

vocabulary beyond their initial few thousand words from reading (Nation, 2001) suggests that how to optimize reading experiences for vocabulary expansion stands as a worthy field of investigation that promises to yield useful educational returns. Given the limited prior research into time intervals between word reencounters from which to draw methodological guidance, findings from chapter 2 inform the present study's overall experimental design, including the manner (and type) of control for potentially confounding factors.

The study addresses a single, basic, issue of both theoretical and practical interest:

> *Will primary-aged EAL children engaged in a school-based RR program learn significantly more vocabulary (statistically and pedagogically) from frequent (or massed) textual reencounters with a novel word of interest (i.e., many encounters during a single RR session) than from encounters dispersed over a broader time span (i.e., RR sessions distributed over several days)?*

To address this question, the study seeks comprehensive answers to the following two Research Questions:

<u>Research Question 1</u>

> *How significant (statistically and pedagogically) are the differences in sums of novel words child readers gain from encountering those words under more, or less, distributed learning conditions? (Does, for example, more learning arise from 'x' reencounters with a word during a single daily RR session than 'x' reencounters with the same word over several daily sessions?)*

The question raises two issues: (a.) how confidently we can presume additional vocabulary gains from relatively spaced encounters with a novel word during RR, and (b.) whether such gains appear pedagogically useful from the perspective of practicing classroom teachers. Because a set offered a uniquely distributed (or massed) presentation of its embedded target vocabulary, the Research Question may be reformulated as follows:

> *How significant (statistically and pedagogically) are the observed differences in the sums of target words participants gained (i.e., total number of known new word-meaning associations) from reading one as opposed to another of the five experimental sets of texts?*

The study addresses Research Question 1 by evaluating the following research hypothesis:

Research hypothesis:

The total of known words from reading one or more sets of texts differs (statistically) from that derived from reading at least one other set of texts.

H_o: The median of known words associated with reading any one set of texts is not significantly different from the median of words gained from reading any other set or sets.

H_a: The median of known words from reading a set of texts is significantly different from the median of words gained from reading at least one other set of experimental sets.

Research Question 2

> *How significant (statistically and pedagogically) are differences in the sums of novel words of the four content word classes (nouns, verbs, adjectives and adverbs) child readers gain from encountering those words under more, or less, distributed learning conditions? (Does, for example, more learning arise from 'x' reencounters with noun 'y' during a single daily RR session, than 'x' reencounters with noun 'y' over several daily sessions?).*

The question builds upon Research Question 1 to ask whether more, or less, spaced learning impacts upon gains of some word classes (the content words: nouns, verbs, adjectives and adverbs) more so than it does others: whether, for example, spaced learning ensures more substantial gains of nouns (N) as opposed to verbs (V), or adjectives (Adj) than adverbs (Adv) and how alternative definitions of 'known word' might qualify findings. Since each set of texts provides a uniquely spaced/massed learning opportunity, the Research Question lends itself to a reformulation analogous to that noted (p. 110) for Research Question 1:

For any lexical class of interest (nouns, verbs, adjectives, or adverbs), did participants learn significantly more words from reading any one set as opposed to another?

The Research Question seeks to establish whether the spacing effect differentially contributes to gains of novel words depending upon the particular lexical class to which those same words belong –whether, for example, the effect proves more 'learning facilitative' for nouns as opposed to verbs, or adjectives as opposed to adverbs etc. This leads on to exploring in detail the significance of any such identified selective effect from a practical, teacher-orientated, perspective.

To addresses Research Question 2 the study tests the following research hypothesis:

<u>Research hypothesis</u>

Research hypothesis: The proportion of known words of a particular lexical class (either noun, or verb, or adjective, or adverb) from reading one set of texts differs (significantly) from the proportion from reading at least one other set.

$\boldsymbol{H_o}$: The sum of known content words children learned (i.e., knew to the relevant standard of knowing) of a particular class from reading any one set of texts is not (statistically) significantly different from that gained from reading any one or more other sets.

$\boldsymbol{H_a}$*:* The sum of known content words children learned (i.e., knew to the relevant standard of knowing) of a particular class from reading any one set of texts is (statistically) significantly different from that gained from reading any one or more other set.

The methodology for addressing the respective hypotheses (Research Questions, 1 and 2) forms the subject of Sections 4.21 and 4.22.

4.3 The study design

The data-elicitation stage (during both the 2008/9 and 2009/10) sessions) took place over four weeks during which each participant worked through the five sets of texts on days and times specified in the reading schedule (see Section 4.19). All reading occurred during regular classroom based RR sessions, each child silently working through their assigned text at their own rate until text completion. Having finished a text, children either went on to read a book from the class library or, if the final text of a set, completed the relevant VSAT test sheet before submitting this to the researcher who would proceed to ask the relevant VSAT verification questions.

Children had access to dictionaries, should they wish to consult them. Participants received notice, however, that some words occurring in their texts might not appear among the entries.

4.4 Group and individual results

This study primarily concerns itself with word-meaning gains of a participant group. Findings concerning individual children receive attention in concluding sections that round off discussions of the two Research Questions (above). These short commentaries aim to bring out notable departures from general observations and help establish the applicability of conclusions to individual cases.

4.5 Words versus word tokens

This study examines word gains from readers' exposures to tokens (i.e. repeated instances of the same words) from working through texts. In Set 1, for example, a single noun (target word) reoccurs 12 times, each occurrence representing a token of that same target word. Should 14 participants successfully learn '*sol*' (to some acknowledged success criteria –see Section 4.23), their gain therefore amounts to 14 tokens of that one target noun as opposed to 14 uniquely different nouns as such. The point has important implications when interpreting study findings. To say that participants learned noun 'x' from reading Set Y means they have mastered, to some standard of knowing, that single noun appearing in Set Y. The total (or maximum) number of discrete words a participant could gain from completing all 5 sets of texts equals 20 (i.e., the sum of the four target words from each set).

4.6 Timing

The dissertation's findings derive from the pooled data from two implementations of the same research methodology[62] –the first occurring during the 2008/2009 academic year (Term 2) and the second during 2009/2010 (Term 1). Two considerations recommended this 'repeat' data gathering process: (1.) More children could take part in the research than would otherwise prove possible (children were keen to participate!) and (2.) The additional student numbers raised the power of relevant statistical tests. Justification for pooling lies in the similar identities of participants in terms of age, prior learning, educational attainment, and the common learning conditions children experienced irrespective of year of participation. Each child, regardless of participation year, followed the same (Year 4) national curriculum; all attended the class of the researcher; each student proved broadly comparable in terms of academic performance (see Section 4.7, below) and each followed essentially the same daily timetable. The slightly lower average age of children in the 2010 study arises from the investigation having taken place in Term 1 of the academic year. The age difference compared to participants in the 2009 limb would not conceivably explain differences in learning outcomes.

4.7 Participants

Participants consisted of 33 children (16 boys and 17 girls) attending Year 4 classes at an international school in Bangkok, Thailand. Each child had achieved SATS scores of English proficiency of between 3b and 3c at the end of the previous academic year, such grades falling in line with the school's age-associated expectations. Participants' home language was Thai (based on parental reports) with English exposure outside of school limited to occasional TV programming and, in some cases, tuition from private tutors. Prior to, and during, the investigation children's primary

[62] Data from the two investigations was treated as if each child had participated on the *same* occasion.

source of English amounted to interaction with classroom teachers. The parents of participants all self-identified as ethnically Thai or Thai/Chinese with the exception of a single Japanese parent (father), who was himself a fluent Thai speaker. No parent claimed a command of English beyond what might loosely correspond to elementary or lower intermediate.[63]

Children's scores from PIPS[64] tests (conducted in June/July of the year prior to participation) proved comparable to those of similarly aged students attending schools in England. The average participant age was 8.6 years (s=0.3) at the time of study completion. The age at which children had first received English instruction fell between 4 to 5 years.

The study focuses on the typical Thai student rather than the Year 4 pupils generally. The following categories of student did not participate in the research project:

1. non-Thai nationals.
2. children receiving SEN support.
3. children from low socio-economic status backgrounds.
4. children who had received lower than a 3c or higher than a 3a in English in the previous year's SAT tests and therefore would receive EAL support, or extension tasks.

Non-Thai children did not participate given the researcher's intention that study conclusions be relevant to the numerically largest population of pupils attending the host institution (Thai nationals) and, hopefully, have applicability to other local international schools delivering instruction in English to a majority Thai L1 student roll. To include SEN students would, similarly, have rendered less robust any conclusions in regard to the target population. This focus upon the hypothetical average Thai pupil also justified the decision to exclude both (a.) particularly able English language learners (defined here as those who entered Year 4 with a SAT grade of 3a or higher), and (b.) children experiencing notable learning difficulties (unassociated with SEN provision), as indicated by inclusion on the EAL register.

[63] This is based on the researcher's assessment during parent–teacher evenings the school held on a termly basis.
[64] Performance Indicators in Primary Schools (see http://www.cemcentre.org/pips/pips)

A final class of non-participants consisted of those from low socio-economic status backgrounds. High school fees meant children attending the host school arrived from privileged expatriate or wealthy Thai families.

In February 2009 and August 2010, eligible Year 4 students then taught by the researcher (n=22, for 2009; n=21, 2010) received a consent form written in Thai and English for the attention of parents/caregivers. The form described the study in broad terms, emphasized the right of children to withdraw from participation at any time point and briefly outlined the VSAT testing process. Others to receive the form included the current Year 4 teachers for whom the research findings would likely prove of particular relevance, and the then head of the Primary EAL Department.

The 33 children selected for the investigation consisted of a random choice from among students enrolled in the researcher's own class who met the eligibility criteria (above). Of this total, five subsequently failed to complete all test sessions and/or the obligatory reading (see Section 5.2 for details). These children's scores did not enter the final data pool from which the study derives its conclusions. Three non-completers were from the 2009/10 limb of the investigation and two from the earlier 2008/09 study (see Table 4.2 below).

2009	**2010**
Boys: 9	Boys: 7
Girls: 8	Girls: 9
Total: 17	Total: 16
Lost to attrition =2	Lost to attrition= 3

Table 4.2: Participant attrition.

4.8 Independent, dependent, and confounding variables

The study involves one independent variable and one dependent variable:

<u>Independent variable (degree of spaced presentation of target word encounters):</u>

Ind. Var. -Length of time measured in number of daily RR sessions (a figure of between 1 and

5 consecutive days) during which participants would encounter all 12 repetitions of the four unique target words embedded in the set of texts they were currently engaged in reading.[65]

Dependent variable:

Dep. Var -The 'amount of learning' arising from the reading experience, with learning gains operationalized as the sum of words known to study participants from having read a particular set of texts (e.g. the sum from reading set 1; the sum from reading set 2, and so on).

The study controls for the following confounding factors:

Factor 1. *Frequency*: The more frequently a reader encounters the unknown word in the text, the greater the chance of acquisition (Horst, 2001; Waring, 2003).

Nature of the control:

All target non-words occur the same number of times (12) within their associated set of texts (see Section 4.12).

Factor 2. *Knowledge of supporting vocabulary*: The probability of a reader acquiring a word meaning from inference strategies depends on the intelligibility of the text as a whole. This intelligibility is itself a function of the ratio of known to unknown lexis (Hill & Thomas, 1988; Hirsh & Nation, 1992).

Nature of the control:

The vocabulary in the experimental texts, apart from target non-words, consists of commonly occurring English words (see Section 4.14).

[65] As noted in Table 4.1, Set 1 presents these repetitions during a single daily RR session, Set 2 over the course of two days, Set 3 over three days, and so on.

Factor 3. *Contextual cues*: Depending upon their type, contextual cues typically prove more supportive of meaning-derivation efforts than others (Ames, 1966; Rankin & Overholser, 1969; Sternberg & Powell, 1983).

Nature of the control:

The 12 occurrences of each non-word occur in the context of contextual clues of comparable informativeness to those in which other target words occur (see Section 4.15).

Factor 4. *Exposure to target words other than in texts*: To the extent words appear in the spoken environment they will tend to reinforce learning gains from reading experiences.

Nature of the control:

The study employs non-words rather than authentic English vocabulary (see Section 4.10).

Factor 5. *Inability to decode lexis*: The more attentional resources diverted to decoding, the less cognitive capacity available to the reader for textual interpretation (Section 2.5.1; see also the *Simple View of Reading* (Gough & Tumner, 1986)).

Nature of the control:

The experimental texts all involve highly decodable vocabulary (see Sections 4.13, 4.14).

Factor 6. *Interest*: In general, readers make fewer gains in word meaning if they engage with texts they find uninteresting or boring.

Nature of the control:

The design and testing of the experimental texts attempted to ensure that participants would find all scripts engrossing (see Section 4.9).

Factor 7. *Cultural familiarity (including background knowledge)*: This represents a well-documented determinant of text comprehensibility (Barnett, 1989; Carrell, 1983; Smith, 1971, 1975). The more familiar the reader with a text's subject matter, the more intelligible the script. That intelligibility, in turn, will impact upon the likely success of GFC efforts.

Nature of the control:

The subject matter of the texts did not reference cultural issues that potentially may cause confusion or misunderstanding (see Section 4.9).

Factor 8. *Part of speech*: The impact of lexical on learnability remains controversial (Nation, 2001), the issue inextricably bound to factors such as imageability and concreteness (Schwanenflugel, 1991). Evidence for a noun-learning advantage (see Elley, 1989; Gentner, 1982), however, suggests a need for controls.

Nature of the control:

Each of the four non-words embedded in a set of texts comprise one of just four lexical classes: noun, verb, adjective, and adverb. The study compares differences in gains of words of each class from reading alternative sets of texts (see Section 4.11).

Factor 9. *Complexity*: Word length (Coles, 1982; Nagy, Anderson, & Herman, 1987), together with grammatical complexity, influences the pleasure children derive from reading experiences (see Section 2.5.1 on the pleasure factor) and the effort they direct towards recovering the textual message.

Nature of the control:

All the experimental texts rate similarly in terms of reading ease, and the embedded non-words, by design, relatively 'short' in terms of constituent orthographic characters and number of syllables (see Sections 4.11, 4.13).

The following sections explain these controls within the context of a broader discussion of the methodology the study employs.

4.9 Materials (controlling for Factors 6 and 7)

Apart from the VSAT test materials (test sheets depicting target words and VSAT states), the study employs 15 specially prepared texts divided up into five sets of reading materials (see Table 4.1

above). The texts consist of liberal adaptations of stories (or themes) from well-known children's authors. The following factors informed the choice of texts:

1. Interest: The texts aimed to engage the reader. Literacy gains from uninteresting texts rarely prove impressive (Krashen, 2004). All texts in the present study consist of narratives for this reason,[66] each chosen after consulting with the school librarian and colleagues familiar with the reading habits of primary-aged Thai children.
2. Familiarity: The text should not recapitulate a storyline already familiar to a child either from prior reading experiences or exposure to other media.
3. Language complexity: All participants should have sufficient reading skills to follow a text's storyline. This called for candidate texts readily comprehensible to even the weakest of the participant readers.
4. Length of texts: Texts could not exceed ~3000 words. This would ensure children had sufficient time to complete a text within specified time limits (see Tables 4.5–4.7 below).[67]
5. Cultural appropriateness: The text should not assume cultural values and beliefs unfamiliar to readers.
6. Importance of illustrations: The current study focuses upon the effects of word learning through reading –specifically, learning that arises from understanding the textual message as opposed to analyzing word-internal cues or drawing interferences from visual displays. Texts retained illustrations only if these did not suggest or imply the meaning of an embedded target non-word (whether the word appeared in that set or another).

[66] The view of colleagues, the librarian, and researcher was that narratives were likely more engaging than expositories.
[67] This assumed an average reading speed of 125 words per minute. The longest text, Text 1, required a full 30 minutes for completion.

4.10 Texts and selection of non-words (controlling for Factors 2, 5, and 9)

The study makes use of non-words incorporated into sets of texts. A non-word, for purposes of the current investigation, is defined as a letter string complying with English phonotactic, graphotactic, and morphological rules but not as yet enjoying 'dictionary status.' The option of employing non-words recommended itself for two reasons: (1.) Participants would not encounter such 'vocabulary' outside of their reading experiences; this allows for the presumption that word gains arise solely from reading experimental texts, and (2.) Non-words obviate the need for pretests to affirm unfamiliarity with target vocabulary, such testing potentially inducing target word learning before children meeting those same words in the experimental texts.

Selecting suitable non-words began with the researcher compiling a list of 70 provisional base-word candidates. Each word ranged from four to six characters (the average word length in English is 4.5 characters) and did not exceed two syllables. To prevent participants attributing meanings to non-words based upon assumed commonality with familiar terms, no candidate word consisted of a homograph or homophone of English or Thai vocabulary[68] or displayed auditory or orthographic similarity to other words on the 70-item list. Decodability was assured (Factor 5) by inviting three native English-speaking teachers to independently evaluate each target word against the reading standard of a hypothetical, weak, Year 4 reader.[69] Words deemed 'difficult' by one or more evaluators were removed from the list. This left a final pool of candidate words (n=67) deemed suitable for inclusion in the present study.

The study intentionally explores word-meaning gains from readers assigning to each target non-word a familiar L1 translation equivalent, as opposed to the more cognitively challenging task of identifying meaning through deductive efforts (see e.g. p.41-42). Two considerations motivated this decision. First, the study would yield insights into the relabeling that explains most vocabulary

[68] The issue involved careful consideration of phonetic distinctions confusing to Thai learners.
[69] All three claimed several years' experience teaching Year 4 Thai children in the host institution.

development among those L2 learners already secure in their mother tongue (Section 2.5.2). Second, the restriction avoids having to control for the cognitive challenge of deriving target word meaning should relabeling prove unavailable. To ensure children possessed the relevant non-word translation equivalents, all non-word candidates consisted of vocabulary either from within the 4,000 most common words in the English language *or*, alternatively, words of particularly high frequency given the language input participants regularly encountered during daily school experiences.

To discourage children 'passing over' words they considered contrived or artificial (this concern had arisen in the discussion following the experiments described in Chapter 3), three teachers rated each non-word for naturalness. This task involved a Likert scale on which each teacher independently assigned the base word forms, along with their inflected variants,[70] to one of four categories: (1.) highly implausible, (2.) somewhat implausible, (3.) plausible, and (4.) very plausible. Words with plausible or above ratings (from all judges) were deemed suitable for inclusion in the experimental texts. Random selection of 20 of these words yielded the specific non-words employed in the current investigation. These appear in Table 4.3 below.

crint lont gos trag harg	parn powl pril rend srep	ned nish wost trop tor	larb sol torg tep garp

Table 4.3: Non-word versions of familiar nouns, adjectives, adverbs, and verbs.

4.11 Assigning class to the non-words (controlling for Factors 8 and 9)

This involved a random number generator that assigned a figure of between 1 and 4 to each non-word candidate. A '1' designated the word as an adjective, '2' an adverb, '3' a noun, and '4' a

[70] It was not sufficient that a base form alone appear natural if the inflected version of that form sounded contrived.

verb. A further round of random number generation (supplying figures 1–5) followed to assign each word to a set of texts (a ‘5’ assigned the word to Set 5, a ‘4’ to Set 4, and so on). The distribution of words (and meanings) appears in Table 4.4 below.

Nouns	Verbs	Adjectives	Adverbs
powl (s) (box) crint (s) (public park) wost (s) (wheel) torg (s) (pond) trop (s) (jacket)	nish (to climb) ned (to shout) tep (to imagine) pril (to argue) lont (to run)	trag (er) (cold) harg (er) (clever) sol (er) (shiny) gos (er) (strong) larb (er) (beautiful)	Parn (ly) (carefully) rend (ly) (quickly) tor (ly) (slowly) garp (ly) (loudly) srep (ly) (happily)

Table 4.4: Non-words by lexical class.

The study employs the following rules to transpose base to derived forms as required by the syntactic environment.

1. Nouns denote plurality by adding an ‘s’ morpheme to their base form (e.g. worsts, torgs etc).
2. Verbs take an ‘s’ (or ‘es’ in the case of ‘nish’) for the third person, an ‘ed’ suffix for the past simple, and ‘ing’ suffix for present participle. For past participles, verbs carry the ‘ed’ suffix.
3. Comparative forms of adjectives form by adding an ‘er’ suffix to the base form.

The study assumes that participants recognize derived forms as instances of the same underlying root word token. Most standard tests of vocabulary size rely on just such a presumption (e.g., Nation’s (2001) Levels tests).

4.12 Placement of non-words into texts (controlling for Factor 1)

Because unknown word meanings limit textual comprehension, no target word lies within 49 word tokens before or after the occurrence of any other target word, including itself. This requirement limits the density of unknown lexis to no more than 2%, a figure that Nation (2000, 2001) considers optimal for learning word/meaning associations. Each of the texts incorporated filler words,[71] as necessary, to raise the sum of intervening vocabulary items between target words placements.

4.13 Adapting the texts (controlling for Factors 2, 5, and 9)

To ensure ready decodability (Factor 5, above), two current Year 4 teachers independently read each text, underlining words deemed sufficiently *challenging* as to interrupt the fluent reading of a hypothetical weak Year 4 student. The extensive classroom experience of those involved offered an assurance that judgments would prove sufficiently objective. The researcher then replaced underlined words with a synonym or a short phrase that preserved, or could substitute for, the meaning of the original term[72] ('caretaker' replaced with 'gardener,' for example, the difference in meaning between the two terms, in this instance, considered irrelevant given the particular text). The reviewer having raised the initial concern then examined the researcher-amended script employing the same test of *challenging* as s/he had with the original text. Further cycles of underlining, assessment, and substitution continued until such time as the reviewer deemed the vocabulary sufficiently decodable to ensure a fluent reading experience.

In order that less accomplished readers would find the texts comprehensible, each script underwent an evaluation using the Kincaid–Flesch grade level readability formula (available on Microsoft Word). The formula yields scores ranging from a possible –3.01 to 12 based on (a.) measures of sentence length, and (b.) syllables per word. A score of '1' indicates a text appropriate for first-grade children and '12' a text suitable for students in the American Grade 12 (ages 17–18), and so forth. The final versions of the experimental texts all scored between 3 and 4, suggestive of suitability for children of chronological ages of around 8 and 9 years, the age of participants in the current investigation. To bring texts within the 3–4 range, those scripts with ratings of 5 or above underwent

[71] That is, words that do not add to the storyline or provide additional details.
[72] In the opinion of the researcher, informed by experience working with Year 4 children.

a process of lexical simplification. This adaptation took two forms: (1.) sentence reduction (i.e. word deletion, or breaking a sentence into two shorter ones) and/or (2.) word replacement (e.g., substituting a word with an alternative of fewer syllables.)[73]

4.14 Moderating vocabulary (controlling for Factors 2 and 5)

Because density of unknown lexis affects comprehension (Nation, 2001), all textual vocabulary other than the target words lies primarily within the 4,000 most common words in the English language. Collectively, these key words comprise approximately 87.6% (Carroll, Davies & Richman) of the vocabulary in a typical script and a somewhat higher percentage of words that serve conversational needs (Schonell, Meddleton, & Shaw, 1956). The strong likelihood that participants would regularly encounter such high frequency words in the school environment strongly implied their meanings would indeed prove familiar. To identify words beyond the 4,000+ range for which substitutions might prove necessary, texts underwent conversion to 'txt' files and inputted into 'Range' (Nation & Heatley, 1994), a software package that draws upon BNC corpus to identify the proportion of vocabulary within various frequency bands of word occurrence. Words beyond the 4,000 ceiling remained in the text only if the researcher considered them (a.) familiar to participants, and (b.) readily decodable. Should a word fail either requirement, then a synonym (a single word or a short, decodable phrase) from the 1–4,000 frequency range served as a replacement.

4.15 Building in contextual clues (controlling for Factor 3)

Among the several factors affecting the likelihood children learn word meanings from RR sessions stands the quality of contextual clues from which meaning derivation proceeds. A word reoccurring in the context[74] of helpful clues will potentially prove somewhat more learnable than one that does not (Ames, 1966; Rankin & Overholder, 1969). The study controls for clue helpfulness as

[73] Number of syllables and sentence length being the variables determinative of Kincaid–Flesch "readability."
[74] The "clue" being a clause in which the target word appeared or an immediately adjoining clause

follows:

1. *Irrespective of the text* and the set in which it appeared, every effort was made to ensure that the particular clues suggestive of target word meaning exhibited similar helpfulness to those in which the same, or other, target words occurred. This effort involved exhaustively assessing the degree of assistance the context afforded (primarily from soliciting the opinions of other members of staff)[75] and then rewriting that clue as necessary to ensure it met the 'comparable helpfulness' standard.

2. *Between any two texts* (any pair from the 15), clues for target words of any lexical class displayed comparable helpfulness. This involved the same consultative process as applied in '1' above.

4.16 Assessing the contextual clues

The following two experiments sought to supply quantitative assurances that contextual clues both within and between texts would prove similar in terms of the assistance they supplied for word-meaning derivation purposes.

Experiment 1

<u>Aim</u>: To determine whether an opportunity sample of English native speakers would supply comparable ratings of helpfulness of 12 clauses in revealing the meaning of a unique non-word embedded in each. The clauses represent a random selection from a single set of texts (Set 4) employed in the current study.

<u>Method</u>: A random sample of 12 clauses containing an embedded target non-word was selected from Texts 1 and 2, of Set 4. Three clauses contained a target noun, three a verb, three an adjective,

[75] Essentially, the process involved copying out sentences in which target words appeared and asking colleagues to rate these for helpfulness.

and a further three an adverb. The clauses were typed out in random order with the English translations of the target word appended in brackets to the end of each. An opportunity sample of eight teachers (English L1, native speakers), plus one non-teaching native English speaker, then rated each clause for helpfulness on a scale of 1 to 12 (ties not permitted, with a '1' designating most helpful and '12' the least). Kendall's coefficient of concordance, a non-parametric test of rater agreement, supplied the quantitative measure of comparability of rankings. Recommendations in Schmidt (1997) served as a basis for interpreting the Kendall statistic ('W'):

0.1 Very weak agreement, None

0.3 Weak agreement, Low

0.5 Moderate agreement, Fair

0.7 Strong agreement, High

0.9 Unusually strong agreement, Very high

<u>Research hypothesis:</u>

Volunteer teachers (n=8) will not rate a sample of 12 clauses similarly.

H_o: Participants displayed evidence of agreement in their rankings (W>0.35; p=0.05).

H_a: Participants did not exhibit evidence of agreement in their rankings (the mean of W<0.35; p=0.05).

<u>Results and conclusion</u>: Results of the Kendall test failed to reveal rating agreement between the eight study participants (average W=0.29), the findings comfortably falling within Schmidt's (1997) range of weak agreement. Data from individual participants indicated no two raters agreed on either the most helpful or least helpful clause, with several correlations indeed proving negative. In only three instances did more than one appraiser (and never more than two) rate the same clause as equally helpful. The null hypothesis is therefore rejected.

Experiment 2:

Aim: To determine whether an opportunity sample of English native speakers would provide similar helpfulness ratings for randomly selected clauses (n=12), each containing both a contextual clue and an embedded target word; clauses were extracted from three different sets of experimental texts.

The experiment seeks to allay concerns that clues in one text could prove more, or less, helpful than those appearing in another.

Method: Twelve clauses, each containing a contextual clue for an embedded target word, were randomly selected from Sets 1, 3, and 5 (four clauses from each set). For each set, one clause contained a target noun, another a verb, the third an adjective, and the fourth an adverb. The clauses were typed out in random order with the English translation of the target word duly appended. A panel of eight English native speaker volunteers, seven of whom taught at the host institution, working independently then ranked each clause according to how helpfully they believed it might assist 'a weak Year 4 reader' derive the embedded word's meaning. Kendall's coefficient of concordance again supplied the statistical measure of agreement, with Schmidt's (1997) guidelines adopted for score interpretation. A follow-up Spearman's rank-order correlation was employed for pairwise comparisons.

$\boldsymbol{H_o}$: Participants displayed agreement in rankings (W>0.35; p=0.05).

$\boldsymbol{H_a}$: Participants did not exhibit signs of agreement in their rankings (W=<0.35; p=0.05).

Results and conclusion: The Kendall's coefficient of concordance revealed only minimal agreement between raters (W=0.32), suggesting that contextual clues displayed comparable helpfulness irrespective of the sets of texts in which they appeared. All eight raters identified uniquely different clauses as the most helpful. Likewise, no two or more raters agreed upon the clause deemed least facilitative, word meaning derivation in mind. In only two instances did any two raters agree on the ranking of the same clause. The null hypothesis is therefore rejected. Findings from Spearman tests revealed only 'very weak' positive correlations, with r^2 values ranging from 0.05-0.25.

4.17 The final evaluation of texts

The final texts (n=15), pictures and illustrations (Elley, 1988) removed, were shared among three colleagues (all experienced Year 4 teachers), each of whom was tasked with singling out clauses that seemed 'unnatural' or 'contrived.' Clauses identified as such were then replaced by the researcher with alternatives that sought to preserve the original meaning, but better satisfied the 'naturalness' test. The original appraiser now evaluated the amended clause employing the same criteria as before. Further rounds of amendment and review followed, as necessary, until the reviewer deemed the clause as sufficiently authentic. A listing of non-words appearing in the final texts along with their occurrences appears in Table 4.5 below.

	Set 1	Set 2	Set 3	Set 4	Set 5
Number of texts in set	1	2	3	4	5
Non-words in texts comprising the set	trop (n) nish (v) harg (adj) rendly (adv)	torg (n) ned (v) trag (adj) torly (adv)	powl (n) tep (v) gos (adj) sreply (adv)	crint (n) pril (v) sol (adj) garply (adv)	wost (n) lont (v) larb (adj) parn (adv)
Number of times each non-word appears in each text of the set	12	6	4	3	2 or 3
Total number of times each non-word appears in the set	12	12	12	12	12

Table 4.5: Features of the selected texts.

4.18 Implementation

Nine days prior to the start of the 2008/2009 investigation, and one week prior to the start of the 2009/2010 repeat study, the researcher conducted a short classroom-based presentation to familiarize children with the research aims and explain what participation entailed. Those expressing interest in taking part received a letter for the attention of parents/caregivers. This communication doubled as a consent form, summarized the research topic and emphasized the child's right to 'opt out' at any point should s/he choose to do so. The text included an invitation to parents/caregivers to contact the researcher should they require further information.

A few days prior to each reading stage commencement date, children undertook a practice session to acquaint them with the testing process. The session involved each child reading an adapted text from the *Oxford Reading Tree* program (Oxford University Press) containing four non-words –a noun, verb, adjective, and adverb each of which appeared between three and four times in the script and in clauses the researcher deemed suggestive of its meaning. Upon completing the text, a child promptly received a VSAT test sheet comparable to those intended for use in the main study. This s/he then worked through unaided before submitting it to the researcher or an assistant (Assistant B, see Chapter 3) who proceeded to ask relevant verification questions, as required (Section 3.3). These pilot tests proved uneventful with all reading and testing concluded within the 35-minute period.

4.19 The reading stage

During the reading stages (2009 and 2010), children read through their allotted texts according to times specified in a reading schedule (Tables 4.6–4.9, below). The schedule identifies children by the group[76] to which each was randomly assigned (1–5) and by an identification number. Numbers in bold designate participants from the 2008/2009 study and those in normal type, the 2009/2010 replication. The schedule served three functions: First, it mitigated possible 'order of treatment' effects –groups of children read the same sets of texts but in a different order to that of other groups. Second, it ensured a manageable number of children presented for testing in the time available during reading

[76] Participants were divided into five groups, each group containing between two and five members.

sessions –achieved by assigning dates on which reading specific sets of texts would occur. Third, and critically, the schedule ensured each set required a unique time-period for completion ranging from between one and five days.

The reading stages (2009 and 2010) extended over three weeks and two days. Each RR session lasted for 25–35 minutes and as comparable to regular RR lessons as the demands of the study permitted. Participants completed their reading in silence after which they selected a class book of their choice or, should they have just finished the last text of a set, went on to work through the relevant VSAT test sheet (see Appendix 3). During RR sessions the researcher and assistants read books/magazines themselves and/or conducted assessments of those presenting completed VSAT test papers.[77] Following an assessment, a child read a book of his/her choice from a selection available in the 'reading corner.' Students not involved in the research (Section 5.3) either read books available in the classroom or the experimental texts, as they preferred.

Student ID numbers (bold =2010)	Mon.	Tues.	Wed.	Thurs.	Fri.
Week 1,Group 1 1,5,9,12,**4, 10,11,17**	Text 4a	Text 4b	Text 4c	Text 4d	Text 1
Week 1, Group 2 2,7,10,15,**1,5,7,15**		Text 1	Text 3a	Text 3b	Text 3c
Week 1, Group 3 3,6,11,16,**6,9,12,14**	Text 5a	Text 5b	Text 5c	Text 5d	Text 5e
Week 1,Group 4 4,8,13,14,**2,3,8,18**	Text 2a	Text 2b	Text 1		
Week 1,Group 5 18,19,17,**13,16**		Text 3a	Text 3b	Text 3c	Text 1

Table 4.6: Week 1 reading schedule.

[77] The assistants were those who participated in the experiments described in Chapter 3. Assistant B conducted VSAT sessions along with the researcher. Assistants A and C served in a consultative fashion.

Student ID numbers (bold =2010)	Mon.	Tues.	Wed.	Thurs.	Fri.
Week 2, Group 1 1,5,9,12, **4, 10,11,17**	5a	5b	5c	5d	5e
Week 2, Group 2 2,7,10,15,**1,5,7,5**	2a	2b			
Week 2, Group 3 3,6,11,16,**6,9,12,14**	1	4a	4b	4c	4d
Week 2, Group 4 4,8,13,14,**2,3,8,18**	4a	4b	4c	4d	
Week 2, Group 5 18,19,17,**13,16**		2a	2b		

Table 4.7: Week 2 reading schedule.

Student ID numbers (bold =2010)	Mon.	Tues.	Wed.	Thurs.	Fri.
Week 3, Group 1 1,5,9,12,**4, 10,11,17**	3a	3b	3c	2a	2b
Week 3, Group 2 2,7,10,15,**1,5,7,15**	5a	5b	5c	5d	5e
Week 3, Group 3 3,6,11,16,**6,9,12,14**	2a	2b			
Week 3, Group 4 4,8,13,14,**2,3,8**,**18**		3a	3b	3c	
Week 3, Group 5 18,19,17,**13,16**	4a	4b	4c	4d	

Table 4.8: Week 3 reading schedule.

Student ID numbers (bold =2010)	Mon.	Tues	Wed.	Thurs.	Fri.
Week 4, Group 1 1,5,9,12,**4, 10,11,17**					
Week 4, Group 2 2,7,10,15,**1,5,7,5**	4a	4b	4c	4d	
Week 4, Group 3 3,6,11,16,**6,9,12,14**	3a	3b	3c		
Week 4, Group 4 4,8,13,14, **2,3,8**,**18**	5a	5b	5c	5d	5e
Week 4, Group 5 18,19,17,**13,16**	5a	5b	5c	5d	5e

Table 4.9: Week 4 reading schedule.

4.20 The VSAT test sheets

To minimize administration time, VSAT testing involved printed sheets on which a child matched words to states by drawing connecting lines (Appendix 3). The test sheets, one for each set of texts, presented 20 words in total: the four embedded target words for the associated Set, plus 16 authentic content words randomly selected from the 6,000 most commonly occurring words listed in the BNC corpus.[78] The 'dictionary' and target non-words appear at random positions on the VSAT test paper. Including authentic vocabulary fulfilled two aims:

1. It concealed target vocabulary among words that may not have appeared in the related set of texts. This would helpfully disincline children from associating VSAT testing exclusively with target non-words and, hopefully, encourage participants to view test sessions as forays into exploring vocabulary more generally.

[78] No attempt was made to ensure an equal number of nouns, verbs, adjectives, and adverbs.

2. Second, the authentic words supplied conversational opportunities to reassure children that word placement in any one state was neither more, or less, 'correct' than placement in another (It seemed important to dispel notions that numerically lower states denoted learning failures, or higher states, successes).

To ensure comparable salience, both target and authentic words were of identical font, color, and text size.

4.21 Addressing Research Question 1 ('*massed* versus *distributed* learning,' word totals)
4.21.1 Methodology for addressing Research Question 1

The study employs two types of analysis to address Research Question 1:

1. A 'general' overview, to identify statistical evidence for a spacing effect that might not present from an analysis restricted to the particular learning gains from reading any one set or another; [79] and
2. A 'pairwise' analysis that looks for statistical differences in word gains from reading one, or the other, set of texts from the 10 possible pairings (i.e. two-set combinations) given the five sets the study employs (i.e., Set 1 versus Set 2; Set 1 versus Set 3 … Set 4 versus Set 5 etc.).

The general analysis aims to reveal indications of a massed or spaced learning advantage based upon the collective learning outcome from the participant group having read all possible combinations of pairs of sets; the 'pairwise' analysis, conversely, seeks to identify among which specific pairs of sets (if any) a significant difference in learning outcomes arises.

1. 'General' analysis[80]

For the 'general' analysis, the study employs a vote-count/sign test procedure. This involves

[79] This can arise when a relatively large (above chance) number of studies return the same result, but the findings in each fail the pre-established test of significance.
[80] For a description of the vote-count method, see Hunter and Schmidt (2004).

identifying both the *sums of pairs of sets* from which the participant group gained more known words from that set offering the more distributed target word presentation (the N+ cases) and the *sums of pairs* from which the participant group gained less (the N- cases). A binomial (one sample) sign test of the N+ and N- cases then evaluates whether the respective sums (i.e. number of N+ and N- values) significantly departs from an equal distribution. To handle tied scores arising from participants failing to learn any words from either text of a pair or from having gained an equal number of words from each, the study adopts an 'apportioning' approach, as opposed to the more conventional practice of ignoring such cases altogether. Should such zero cases (i.e. tied scores) amount to two or more, they become 'shared' equally among the N+ and N- observations. In the event of an odd number of zeros, the extra case was allocated to the N+ or N- total –whichever of the two would most likely preserve the null hypothesis presumption.

H_o: $\pi = 0.5$ (no evidence for a spacing effect)

H_a: $\pi > 0.5$ (a spacing effect is presumed)

In instances where the sign test returned a significant value ($p<0.05$), the direction of the learning advantage (i.e. whether arising from the more massed or spaced target word encounters) was presumed from the respective number of N+ and N- cases. Should the N+ cases significantly ($p<0.05$) exceed the sum of N- the learning advantage was attributed to more spaced (or longer intervals) between target word presentations. Conversely, if the total of N- cases exceeded N+ cases, the study assumes more learning from sets offering the relatively massed learning condition.

2. 'Pairwise' set-to-set comparisons

Aim: To identify cases of significant differences in the number of target words readers knew from reading one set of texts compared to gains from reading an alternative set.

Research hypotheses:

H_o: The sum of words known to participants (as a group) from reading any one set of texts does not significantly differ from the sum gained from having read any one (or more) other sets of the five.

H_a: The sum of words known to participants (as a group) from reading any one set of texts does significantly differ from the sum gained from having read at least one other set of experimental sets.

Method: Statistical procedure

Given the within-subjects form of the investigation (the same participants read each set of texts from which known word totals derive), the following statistical techniques recommended themselves:

a. Repeated-measures ANOVA (parametric) followed by multiple cross comparisons wherever the F statistic indicates a significant difference in means.
b. The Friedman test (non-parametric) for differences in treatment medians, with follow up Wilcoxon post-tests, or sign tests, in the event of a significant Chi square finding.

The study adopts the Friedman-based analysis as the more likely to supply robust data. This preference stems from the anticipated skewed dispersal of VSAT scores (a preponderance of words in State 1) violating the normality condition that is a precondition for robust ANOVA analyses. For post-tests, following on from a significant Friedman's finding, the study employs the 'two dependent samples sign test' for pairwise comparisons.[81] The test amounts to a simple, distribution free, indicator of the likelihood that the samples derive from different populations.

To minimize type 1 errors, the study follows the Hollander and Wolfe (1999) recommendation that all zero scores be evenly distributed among the N+ and N- observations –the same practice as for the general analysis noted earlier (see p.133). This apportioning acknowledges that zero scores

[81] The study does not employ Wilcoxon post-tests given the many anticipated tied scores (the likelihood of the same participant learning the same number of words from two or more sets of texts), which breach the Wilcoxon presumption of continuity.[81]

indirectly support the null hypothesis conjecture and represents a compromise position between what Geyer (2013) calls the "zero fudge" (ignoring zeros altogether) and the alternative but conservative (Hollander & Wolfe, 1999) option of assigning all zeros the value (either + or -) least supportive of *Ho; i.e.*, the H_o: $\pi+= 0.5$ position.

All VSAT testing took place promptly after a child completed a set of texts,[82] the data supplying totals of words occupying each VSAT state together with the all-important 'known word' sums (see Section 4.23).

Pedagogical significance

The second limb of Question 1 moves beyond statistical significance to identify the pedagogical usefulness of additional word gains from more spaced encounters. To address this issue, the study expresses statistically significant additional word gains as:

1. The sums and averages of target words children came to know (of a particular lexical class of interest) from reading one set of a pair relative to those gained from reading the other.

2. The proportion of those additional words children knew to the various lexical competences represented by the VSAT knowledge states –whether, for example, a significant difference is more reasonably attributed to additional words disproportionately occupying certain states as opposed to others (e.g. State 6 than State 5, or State 5 than State 4), as opposed to a dispersion of known words among VSAT states generally. For determining whether sums in the same state among different sets proved significantly different the study employs binomial sign tests, adopting the same apportioning of zeros as noted above (p.135).

[82] This provided data of the total number of words (and lexical class of words) each of the 28 participants gained from reading each set.

To explore the pedagogical 'usefulness' of gains from a teacher's perspective, the study estimates and compares differences in known word gains over an academic year, assuming RR sessions in which children read texts conforming to the specifications of one, or another, of those that the current study employs. For interpreting differences in known sums, the study introduces the construct of standard measures (SMs), where an SM represents the hypothesized sum of words that children comparable to participants in the present investigation might have learned during a single academic year from unadapted reading materials –i.e. the regular texts engaged with during RR sessions. The one-year period recommended itself as a 'traditional' duration of academic programs.

The study estimates SMs through the following three-step procedure based loosely on Nagy et al. (1985):

Step 1. Identifying the sum of novel words a child will likely encounter during RR sessions from unadapted (regular) texts over the course of one academic year.

This involves establishing figures for:

1. the number of RR sessions per year,
2. the duration of those sessions,
3. the child's average reading speed, and
4. the proportion of the total words read that are likely to prove unknown. A rough estimation is 1 to 2 percent assuming teacher supply children with RR tests with an optimal number of novel words for purposes of vocabulary growth (Hu & Nation, 2000).

Step 2. Estimating the sum of novel words children actually gained of the total novel words encountered per year. Such gains (i.e. the SM) amount to the product of the probability of word gain from a single encounter (assuming conventional unadapted RR texts –See Section 5.19.1 for details), and the figure derived in Step one.

Step 3. Estimating the proportion of novel words a child might gain from one year of RR sessions having read adapted texts of the sort the current study employs. The figure is determined in the manner described in Step 2, except that the probability of learning now derives from children's word gains having read the researcher designed sets of texts, Sets 1 to 5.

Step 4. Expressing the sum of words derived from applying Step 3 as a proportion of the SM (i.e., the sum derived from applying Step 2).

With the SM established, the study compares the excess sums of known words from reading RR materials designed to the specifications of one or another of the experimental sets of texts, against this presumed (SM) annual gain total.

4.22 **Addressing Research Question 2 ('the issue of *massed* versus *distributed* learning at the level of word class')**

4.22.1 Methodology for addressing Research Question 2

The study employs the same two-stage analysis adopted for exploring gross known word sums: (1.) A 'general' analysis (this draws upon the vote-count/sign test procedure -see above) to identify an effect that might not reveal itself in statistically significant gains in comparisons of word gains associated with reading one set texts as opposed to another, and (2.) set-to-set, 'pairwise' analyses of differences in known word totals of a class from reading one set of a possible pair.

1. 'General analysis.' A spaced learning advantage is presumed if the number of pairs of sets (of the possible 10) from which children gained an excess of words of any one lexical class (N+ observations) significantly ($p<0.05$) exceeded the number from which they gained less (N- observations), the null hypothesis retained, otherwise. The null and research hypotheses are those the study employs in the analysis of gross sums, namely:

    H_o: $\pi = 0.5$

    H_a: $\pi > 0.5$ (The null hypothesis is rejected if $p<0.05$).

Tied scores receive treatment in the manner described for gross known word sums (Section 4.21.2) and undergo the same allocation between the N+ and N- data sets.

2. ‘Pairwise’ analysis

Aim: To determine the significance (if any) of differences in the number of known content (target) words of a specific type (noun, verb, adjective, adverb) from participants having read one set of texts relative to those from reading another. The raw data for these comparisons derives from the same VSAT supplied records of word-to-state assignments employed to address Research Question 1.

Research hypothesis: The proportion of known words of a particular lexical class (noun, verb, adjective, or adverb) from reading one set of texts differs (significantly) from the proportion observed from reading at least one other set.

$\boldsymbol{H_o}$: The sum of known content words of a particular class from reading any one set of texts is not (statistically) significantly different ($p>0.05$) from the sum gained from reading at least one other set.

$\boldsymbol{H_a}$: The sum of known content words of a particular class from reading any one set of texts is (statistically) significantly different ($p<0.05$) from the sum gained from reading one, or more, other sets.

Method: Statistical procedure

The approach involves a between-sets analysis employing Cochran’s Q, a non-parametric measure for equality of proportions with dichotomous data. The test recommends itself given the binary quality of VSAT derived data: a student either knows a word, or s/he does not. The study employs the following two-step procedure (Harris, 2001) to identify significant differences in the proportions:

a. Computing the Q statistic and Q/df ratio (where df = degrees of freedom); The Q statistic reveals significant differences in the proportion of words of the same lexical class from reading any one set or sets as opposed to any other.

b. Should either (1.) Q prove significant or (2.) Q prove non-significant but the ratio of Q and the degrees of freedom (i.e., Q/df) is < 1, conducting post-hoc McNemar tests to determine between which combinations of sets any significant difference(s) lies.

Pedagogical significance

The study explores the issue of pedagogical significance in an analogous fashion to that adopted to investigate the same topic in the context of Research Question 1. The methodology requires computing differences in:

1. The sums of words gained (of a particular lexical class of interest) from reading one set of texts a possible pair over and above those children gained from reading the alternative set.
2. The proportions of additional words of a class children knew to the standard of *knowing* that an individual VSAT state, or combination of states (See Section 4.23), attempts to capture (e.g. Did, for example, children learn as many words of a particular type -whether noun, verb, adjective or adverb- in VSAT *State 5* from set 1 as opposed to Set 4, or as many from Set 2 as opposed to Set 3?). The study applies McNemar tests to identify significant differences in learned word sums.

To explore pedagogical significance further, the study adopts the construct of SMs described previously in the discussion of gross known word totals, but now applies this purely to comparing sums of known words of each lexical class of interest (nouns, verbs, adjectives, and adverbs). Drawing largely upon Nagy et al. (1985, 1987), this comparison involves two steps:

Step 1. Apportioning the presumed yearly gross sum of learned words from *unadapted* reading materials among the lexical classes of interest in the current study: namely nouns, verbs, adjectives, or adverbs. This apportioning, itself, involves the following:

1. Identifying a *relevant proportion* of known sums for each lexical class

The study assumes that the percentage of words of a class of interest corresponds to the proportion of that class among the 4,000 most commonly occurring words in the English language as reported in the BNC. These words comprise a substantial 87.9% of the vocabulary appearing in any text (Nation, 2001). If, for example, verbs were to comprise 25% of that sum, then the relevant proportion amounts to 0.25; if words of this class comprised 34% then the proportion rises somewhat to 0.34, and so on.

2. Applying the *relevant proportion* to deduce SMs.

With the relevant proportion established, the sum of known words of any one lexical class of the total gains from unadapted texts is presumed to equal the product of the relevant proportion for that class, and the gross (undifferentiated) sum of words gained during the one year of RR sessions. If we assume, for example, the relevant proportion for nouns amounts to 0.5, and the gross sum of learned words undifferentiated by lexical class totals 100, then the assumption leads us to suppose that 50% of the learned words belong to the noun category. The SM figure for nouns in this case amounts to 50 (i.e. 0.5 x100).

Step 2. Expressing gains in SM form.

The final figures from Step 1 serve as the SMs against which to evaluate differences in known words (by class) from reading one or another set of experimental texts the current study employs. Children's word gains (i.e., the sum of additional words) from reading the set of a pair associated with the most learning, expressed in SM terms, equals the sum of those gains divided by the SM for the lexical class of interest.

4.23 What is a *known* word?

How much word learning arises from RR depends upon what amounts to a known word, the more the stringent test, the less learning we would likely observe per reading experience. In the absence of agreement, and an ambiguity among linguists regarding what 'knowing' necessarily implies (See e.g. Nation, 2000), the study analyzes each Research Question from the perspective of three alternative notions of what 'to know a word' might reasonably entail. These notions correspond to a particular lexical competence, or group of competencies, represented by one or more VSAT states (see Section 3.3). Specifically:

1. **Definition 1** of known words (most strict/least inclusive): A known word is one that occupies VSAT State 6.
2. **Definition 2** of known words (less strict/more inclusive): A known word is one that occupies VSAT State 6 or 5.
3. **Definition 3** of known words (least strict/most inclusive): A known word is one that occupies VSAT States 6, 5, or 4.

Descriptions of each state appear in Chapter 4, along with details of the VSAT administration procedure.

4.24 Statistical significance

Following well-established convention, a test yields a statistically significant finding if the likelihood of an occurrence amounts to less than 0.05 (5%), assuming the null hypothesis conjecture holds true. The current study accordingly adopts this traditional $p<0.05$ value for interpreting both Friedman and follow up sign test findings. For Cochran Q tests, given their low power to detect heterogeneity, the study accepts Heneghan and Badenoch's (2006) recommendation that alpha be set at 0.1, thereby addressing a concern that the test may supply p-values that miss significance despite that McNemar analyses on the same pairs of data may indeed reveal significant difference findings. Should a Cochran's Q supply a p-value lesser than 0.1, the study adopts the conventional alpha value of 0.05 for the McNemar tests that follow.

The study does not employ Bonferroni correction with post-hoc tests since the investigation does not concern itself with family-wise error (the likelihood of one or more type 1 errors from a test of multiple hypotheses) but rather with significant differences in learning from reading one or another set (of the texts) of a pair. All statistical calculations were performed with SPSS software, versions 17 and 20.

4.25 Anticipated outcomes

Vocabulary acquisition proceeds incrementally from multiple exposures to a word in different contextual settings, each exposure potentially consolidating and/or building upon preexisting knowledge. Because the experimental texts incorporate contextual clues, children will have ample opportunities to acquire the meanings of the embedded target words in each set of texts they complete during the course of the investigation. The findings from spacing effect research suggest that children will display more, or less, word learning depending upon the set of texts they have completed, with more learning arising from reading Sets 5 and 4 than Sets 1 or 2. For both sums of *known* words overall, and of those of a particular lexical class of interest, how impressive the observed differences in word gains from reading one, or another, set of texts may depend upon the notion of *known* word one chooses to acknowledge.

4.26 Summary

Few studies have explored how successfully children acquire denotational knowledge from reading experiences, and none has related that learning to more or less distributed word presentations. The current chapter set out to describe, and justify, a practical methodology for exploring and quantifying the effects of spaced learning from RR in an authentic school setting. The discussion defined the particular student population of interest, described the design and construction of experimental sets of texts, identified confounding factors (and relevant controls), and reported upon the particular selection of participants. The chapter concluded by describing the process of data analysis along with the relevant statistical tests from which the study seeks credible responses to the Research Questions.

Chapter 5

Results and Findings

5.1 Introduction

The current chapter presents and interprets the results of the current study in an effort to derive comprehensive answers to the two Research Questions of interest (Section 4.2.1). Following a brief commentary on the final data pool and manner of data presentation, the chapter divides into three Parts. Part 1 (pp. 147-163) reports statistical findings, employing graphical displays, tables and accompanying text to identify significant differences in word gains from reading alternative sets (of texts) of a possible pair. The discussion addresses pedagogical significance by references to (a.) actual sums of *known*[83] words, (b.) averages of *known* word gains, and (c.) totals of *known* words differentiated by lexical class. Part 2 (pp.164-227) elaborates upon Part 1 findings by exploring pedagogical importance in rather more detail, focusing upon those pairings of sets of texts (revealed in Part 1) from which statistically significant differences in learning arose. The discussion reports upon (a.) the difference in the sums of word gains –notably points gleaned the dispersion of *known* words among the VSAT states, and (b.) what those gains imply in terms of lexical competencies.

Part 3 (pp. 227-252) builds upon both Parts 1 and 2 to shed light on the pedagogical importance of spaced learning from a practical, teacher orientated, perspective. The discussion presents estimations of RR word gains, both differentiated and undifferentiated by word class, presumed to arise over the course of a hypothetical academic year supposing children were to read texts designed to the specifications of one or another experimental set.[84] This analysis places spaced learning into an applied context, providing an indication of word gains over a time period meaningful to classroom instructors. The discussion draws heavily upon hypothetical estimates of *known* word totals assuming

[83] 'Known' is italicized in the present chapter as a reminder that the term has three alternative definitions in the context of the current study.

[84] This involves estimating, for example, how many words children would gain were they to read texts designed to the specification of set 1 as opposed to set 2, or set 3 as opposed to set 5, for example.

children were to have read unadapted RR texts over the same one-year duration (SMs).[85] The final chapter, Chapter 6, discusses the theoretical and practical relevance of the current research to teaching practices and child vocabulary acquisition in general.

5.2 The final data pool

The study recruited an initial 33 children from two Year 4 classes, each taught by the author over the course of the 2008/2009 and 2009/2010 academic years. Two participants left the school during the investigation (one during the first academic year and one during the second), while a further two failed to attend several classes due to concerns over H1N1 (Term 2, 2009). An additional child could not complete VSAT tests at one or more specified time points due to ill-health or conflicting commitments (Term 1, 2010). Children who resumed schooling (n=3) continued to read the experimental texts and underwent testing along with the other participants. Their scores, however, did not contribute to the data pool from which the study derives its conclusions. In total, 28 children (33 minus 5) completed the entire reading and testing program in accordance with the study design.

5.3 Notes on data presentation

The study reports findings in regard to both gross *known* word sums and sums by *lexical class* (noun, verb, adjective, adverb) for each definition of *known* word defined in Section 4.23, namely: (1.) words VSAT testing assigned to State 6; (2.) words occupying VSAT States 6+5; and (3.) words occupying VSAT States 6+5+4. For the definition of interest, the discussion begins with an overview of the raw data involving a commentary on differences in *known* word totals, averages, and percentage gains from the participant group reading one or another set of texts of a possible pair. Following this, the study reports statistical findings to establish the probability that disparities in *known* word sums (from reading a set of a pair) indeed arise from the factor of interest –the time differences over which each set of texts presented its embodied target words (Research Question 1). The final section of each

[85] The point being that if a child gains 100 words extra from distributed learning, then this is rather more meaningful if it amounts to, say, 50% of the total words the typical child gains per year, as opposed to 2%.

analysis asks whether children gained significantly more words of the same lexical class (Noun, Verb, Adjective or Adverb) from reading one set of texts relative to gains from having read another –the concern of Research Question 2.

Part 1 (Statistical findings)

5.4 *Known* words as those in VSAT State 6

5.4.1 State 6 raw data and statistical analyses

Table 5.1 and Figure 5.1 display participants' *known* word totals from having read each set of texts (1–5), where a *known* word represents any that VSAT testing assigned to State 6. Occupancy of this state requires that two conditions apply:

1. The participant could supply that target word in a syntactically and semantically well-formed sentence (see Section 3.5), and

2. The word is one for which the test taker provided a 'suitable' synonym (Section 3.5).

The sums of *known* words for each individual child appears in Appendix 2a.

General comments

The *known* target word sums for the participant group (the aggregate of children's individual scores) ranged from 31 out of 112 (highest: Set 5), to 24 (lowest: Set 3). The single most learnable class emerged as nouns, with children gaining more words of this type than of any other irrespective of set of texts. Adverbs proved the generally least learnable, with participants as a group gaining fewer words of this category than of any other from all sets of texts aside from Set 3 from which verb gains proved marginally lower (Table 5.1). The sum of *known* adjectives exceeded that for verbs from reading Sets 1 to 4, while verbs outnumbered adjectives from reading Set 5. The mean word gain per child ranged from a high of 1.11 (Set 5; sd= 0.63) to a low of 0.86 (Set 3; sd= 0.93) –a difference of 0.25 words. The percentage of target words the participant group gained from a single set of texts extended from a low of 21% for Set 3 to a high of 28% for Set 5.

Set	Noun	Verb	Adj.	Adv.	Average number of words learned by each participant	Percentage learned of total (112)
1	14	4	6	3	(av=0.96); sd. 0.69	24
2	12	5	5	3	(av=0.89); sd. 0.74	22
3	10	2	9	3	(av=0.86); sd. 0.93	21
4	12	5	7	3	(av=0.96); sd 0.79	24
5	15	6	5	5	(av=1.11); sd 0.63	28
	63	22	32	17		

Table 5.1: *Known* words, by set and lexical class (av=average; sd=standard deviation), State 6.

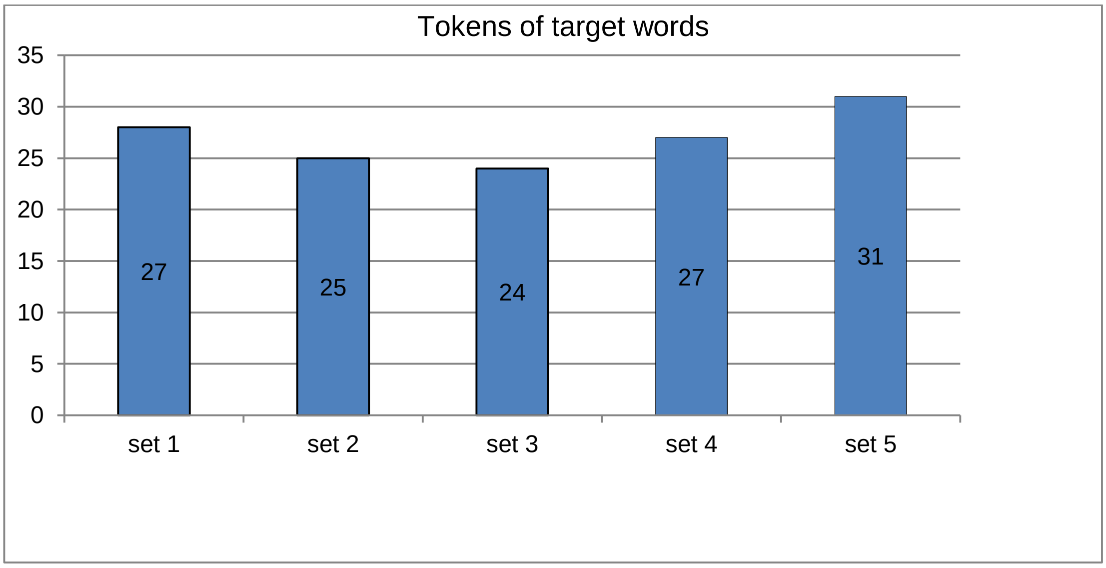


Figure 5.1: Totals of target word tokens gained, by set, State 6.

5.4.2 Research Question 1 (State 6): *How substantial, statistically, are differences in the number of known words from reading one set of texts as opposed to another?*

Statistical Finding:

From the 10 possible pairings of sets, children gained additional words from the more spaced set in 6 (N+) cases, and less in 3 (N-), with one case of tied score, the latter assigned to N-. The vote-count/sign test returned a p-value of 0.75, a value consistent with the null hypothesis (Section 4.21.1) of no additional word gains arising from spaced learning opportunities. A Friedman's ANOVA conducted to uncover possible significant differences in *known* word totals from children having read one or the other set of a possible pair (the 'pairwise' analysis) similarly failed to indicate a spaced learning advantage (df=4, chi-square=1.89, p=0.76). This observation holds true at both alpha =0.05 and 0.1. No follow up sign tests proved warranted.

Conclusion:

The totals of target words participants *knew* to the standard of VSAT State 6 do not indicate significantly more word gains from the set of texts of a pair providing the more distributed target word encounters. Both the 'general' and 'pairwise' null hypotheses (see Section 4.21.1) remain intact.

5.4.3 Research Question 2 (State 6): *How substantial, statistically, are differences in the sums of nouns, verbs, adjectives or adverbs from reading any one set as opposed to another?*

Statistical Finding:

The vote-count/sign test on the N+ and N- observations returned p-values ranging from 0.34 to 1.000 (two-tailed),[86] affording no evidence for a 'general' spaced learning effect on the *known* sums of any word class. Cochran's Q tests to identify possible 'pairwise' differences in word proportions (Figures 5.2–5.6) indicated that the participant group failed to gain significantly more words of any type from reading any set of a possible pairing. Data for individual participant's word gains by lexical class appears in Appendix 2a.

[86] Nouns (p=1.00); verbs (p=0.34); adjectives (p=1.00); adverbs (p=0.34).

N	28
Cochran's Q	2.41
Df	4
Asymp. sig	0.66

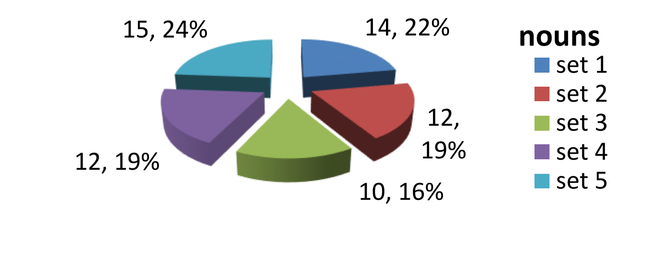


Figure 5.2: Percentage of nouns assigned to State 6, by set of texts.

N	28
Cochran's Q	2.71
Df	4
Asymp. sig	0.61

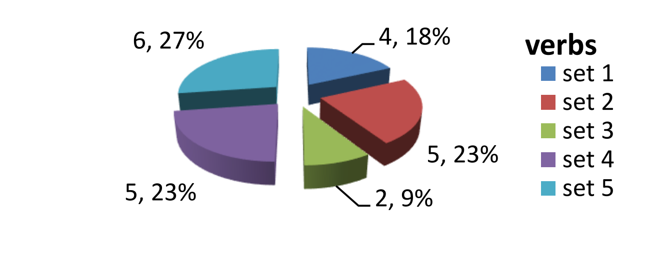


Figure 5.3: Percentage of verbs assigned to State 6, by set of texts.

N	28
Cochran's Q	2.15
Df	4
Asymp. sig	0.71

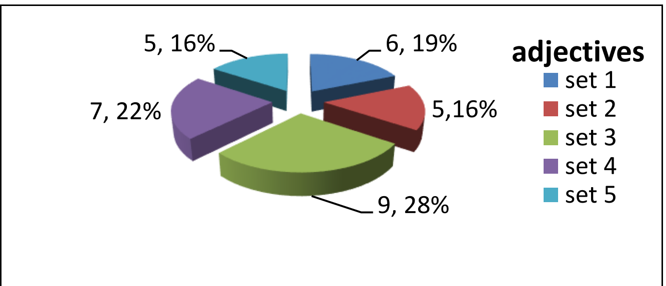


Figure 5.4: Percentage of adjectives assigned to State 6, by set of texts.

N	28
Cochran's Q	1.39
Df	4
Asymp. sig	0.85

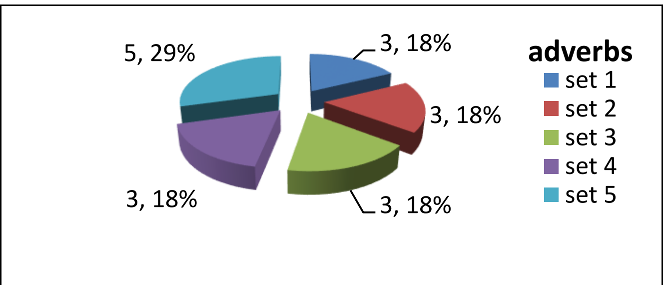


Figure 5.5: Percentage of adverbs assigned to State 6, by set of texts.

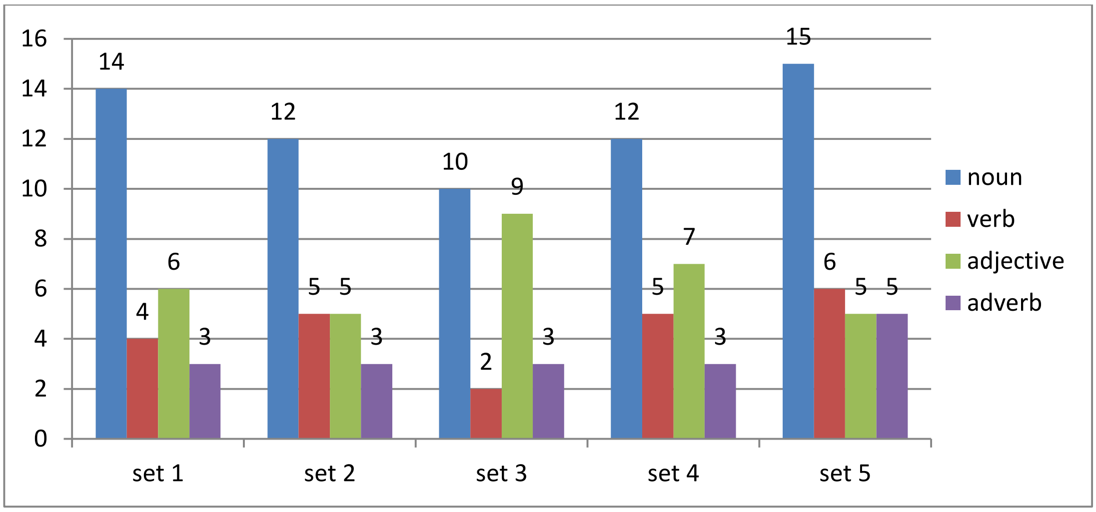


Figure 5.6: The proportion of tokens assigned to VSAT State 6 by set and class.

<u>Conclusion</u>:

The participant group did not gain significantly more words of any one class from encountering those words under more distributed learning conditions, with *known* words defined as those exclusively occupying VSAT State 6. The 'general' and 'pairwise' null hypotheses (i.e. no spacing effect learning advantage) are retained.

5.4.4 Notes on *Individual* learning outcomes (*known* words, State 6)

The total *known* words for single participants ranged from 0 to 3, though scores displayed some variability across and within sets: Participant 19, for example, gained zero words from Sets 1, 2, and 3 (ranking joint last) yet 2 words each from Sets 4 and 5 (ranking 2nd in each instance); participant 15 *knew* no words to the State 6 criterion from reading Sets 5 and 1, ranking last among peers, yet 1 word each from Sets 2, 3, and 4 (ranking respectively 2nd, 3rd and 3rd). No-scorers (i.e. children who failed to learn any target words) amounted to 7 for Set 1, 9 for Set 2, and 12, 8 and 4 for Sets 3, 4, and 5 respectively. Only 5 children (participants: 8, 13, 19, 25 and 27) exhibited a consistent pattern of equal or additional word gains from the set of a pair offering more distributed learning opportunity. No child (see Appendix 2a), however, displayed the reverse condition –i.e. less or equal gains from a set providing a more massed learning experience. The highest proportion of *known* words achieved from a set of texts amounted to 75%, this from participants 3 and 18 from having read Set 3, and participant 5 from reading Set 4.

5.5 *Known* words as those in VSAT States 6+5

5.5.1 States 6+5 raw data and statistical analyses

Table 5.2 and Figure 5.7, below, present statistical summaries of *known* word data where *known* signifies word occupancy of VSAT States 6 or 5. Words in these states consist of:

1. Target words that a participant could supply in a grammatically and semantically well-formed clause (see Section 3.5), and/or
2. Any word for which the test taker could produce a native-like synonym.

For a record of each child's individual learning outcomes see Appendix 2b.

<u>General comments</u>

Including VSAT State 5 words in the *known* category resulted in a predictable increase in learned word totals. From reading Set 1, participants gained an additional 24 words (an 89% increase from the sum associated with the State 6 test of *knowing*), for Set 2, gains amount to 24 extra words (an increase of 96%), and for Sets 3, 4 and 5 gains of 31 (129%), 38 (140%) and 40 (129%) words respectively. The highest *known* word sum from reading a set of texts came to 71 (this from the participant group having reading Set 5) and the lowest, 49 (from reading Set 2), a difference of 22. Nouns once again proved the most learnable class, and adverbs the least. Totals of learned verbs exceeded those for adjectives from reading Set 3, while adjectives proved the more learnable from Sets 1, 2 and 4. The average of word gains by set ranged from a high of 2.54 (Set 5; sd= 0.79), to a low of 1.75 (Set 2; sd=0.80) – a difference of 0.79 words. The percentage of words learned per set (from a possible maximum score of 112) ranged from 44% (Set 2) to 63% (Set 5).

Set	Noun	Verbs	Adj.	Adv.	Average number of words learned by each participant	Percentage Learned of total
1	22	10	13	6	(av=1.82); sd. 0.86	45.5
2	17	11	13	8	(av=1.75); sd. 0.80	44.0
3	19	15	11	10	(av=1.96); sd. 1.07	49.1
4	22	16	17	10	(av=2.32); sd. 0.86	58.0
5	27	17	17	10	(av=2.54); sd. 0.79	63.3
	107	69	71	44		

Table 5.2: Known words (States 6+5), by set and lexical class (av=average; sd=standard deviation).

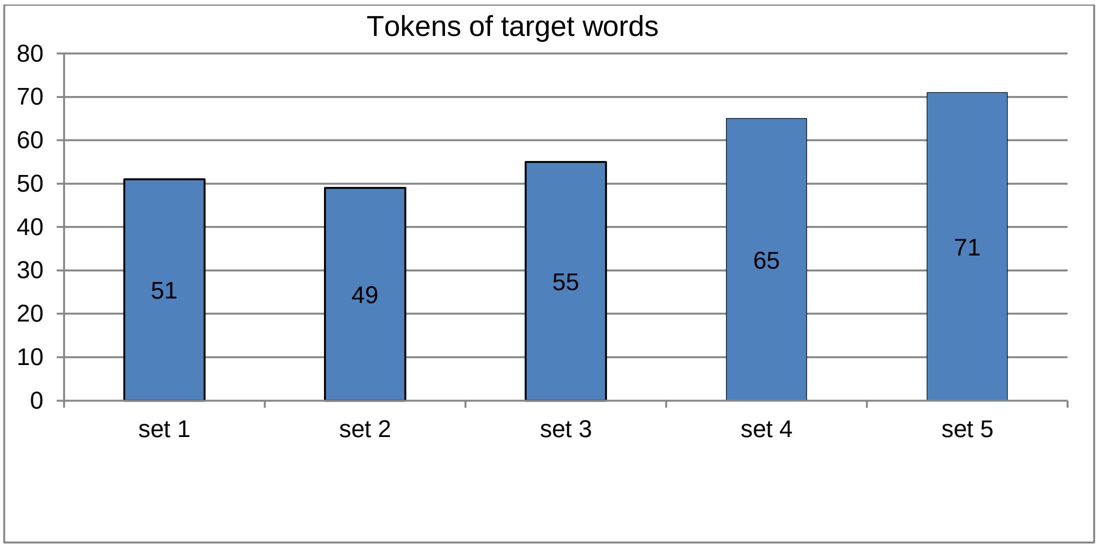


Figure 5.7: Totals of target words gained by set, States 6+5.

5.5.2 Research Question 1 (States 6+5): *How substantial, statistically, are differences in the number of 'known' words from reading any one set as opposed to another?*

<u>Statistical Finding</u>:

The vote-count/sign test for a spaced learning effect supplies 'strong' evidence for null hypothesis rejection, the participant group having gained additional words from the more spaced set of a possible pair in 9 out of 10 instances ($p=0.021$; two-tailed). A Friedman's test ($p < 0.002$, $df=4$, Chi-square =17.34) and follow up sign tests for 'pairwise' effects indicated a significant difference in word totals in four instances: (a.) from participants having read Sets 5 and 1 (p=0.004); (b.) from reading Sets 5 and 2 (p=0.003); (c.) from Sets 5 and 3 (p=.007); and (d.) from reading Sets 4 and 2 (p=0.019). A further sign test, with zeros apportioned (See Section 4.21.1), revealed significant differences (two-tailed) remained for the following set pairings (Sets 5 and 1, p=0.01; Sets 5 and 2, p=0.01; Sets 5 and 3, p=0.01; Sets 4 and 2, p=0.04).

Conclusion:

Both the vote-count/sign test, and sign testing for 'pairwise' comparisons returned significant p-values (in for cases) consistent with the research hypothesis (*Ha*) of children having gained more target word encounters from spaced word presentations. Neither the 'general' or 'pairwise' tests support the null hypotheses conjecture (Section 4.21.1).

5.5.3 Research Question 2 (States 6+5): *How substantial, statistically, are differences in the sums of nouns, verbs, adjectives or adverbs from reading any one set as opposed to another?*

Statistical Finding:

The vote-count/sign test supplied p-values from p=0.002 to 0.34 returning a significant difference in the median of N+ and N- cases for a single lexical class, verbs (p=0.002; two-tailed); the findings imply that the participant group gained more words of this category from sets presenting target vocabulary over more spaced presentation intervals. The findings for both adjectives (p=0.34) and nouns (p=0.34) miss significance comfortably, as does the p-value for adverbs (p=0.10). Results from Cochran's Q tests and follow up McNemar ('pairwise') tests, however, indicated significant differences in *known* word sums in two instances: (a.) in the proportion of nouns from children having read Sets 2 and 5, (p=0.002), and (b.) in the proportion of nouns from Sets 3 and 5 (p=0.008). No such significant differences emerged for *known* verbs, adjectives or adverbs from children having read any set of a pair.

N	28
Cochran's Q	13.000
Df	4
Asymp. sig	0.01

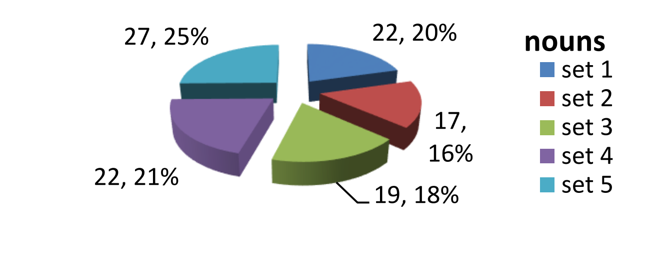


Figure 5.8: Percentage of nouns assigned to States 6+5, by set of texts.

N	28
Cochran's Q	5.62
Df	4
Asymp. sig	0.23

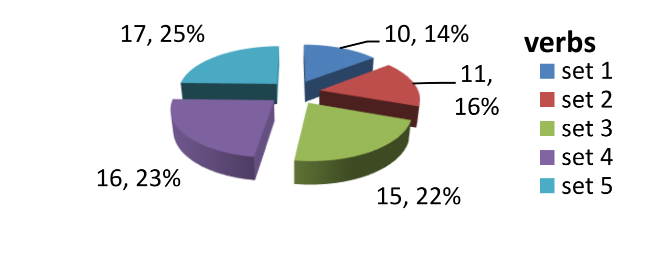


Figure 5.9: Percentage of verbs assigned to States 6+5, by set of texts.

N	28
Cochran's Q	4.06
Df	4
Asymp. sig	0.40

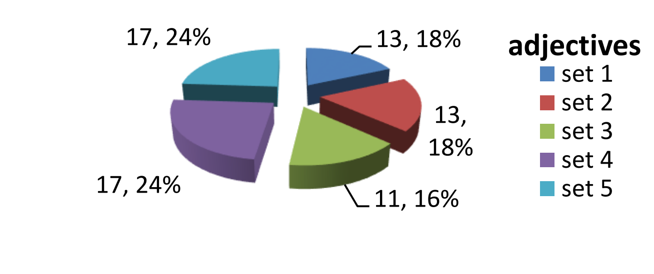


Figure 5.10: Percentage of adjectives assigned to States 6+5, by set of texts.

N	28
Cochran's Q	2.42
Df	4
Asymp. sig	0.66

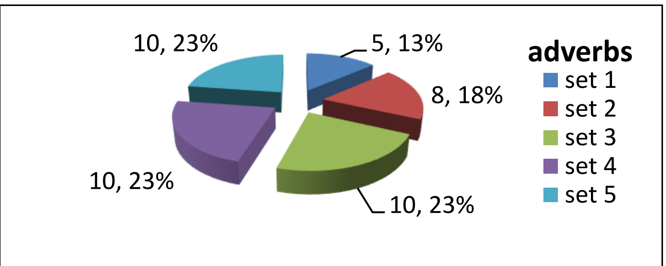


Figure 5.11: Percentage of adverbs assigned to States 6+5, by set of texts.

<u>Conclusion</u>:

Findings from the vote-count/sign test refute the null hypothesis (Section 4.22.1) conjecture, and lend support to the research hypothesis (H_a) that extensions to the time over which readers encountered target verbs (but only verbs) indeed contributed in some manor to the more impressive learning. McNemar tests for a 'pairwise' effect identified two cases (Sets 2 and 5, and Sets 3 and 5) from which the participant group successfully gained additional nouns from the set presenting the more distributed learning condition –this despite the 'general' test having failed to do so. No statistically significant differences emerged in the proportions of *known* verbs (Q=5.62; df=4, p=0.22) albeit the vote-count/sign test returned a highly significant finding (p=0.002) –see Section 5.12.15 for a discussion of this apparent anomaly. Both the 'general' and 'pairwise' null hypotheses are rejected.

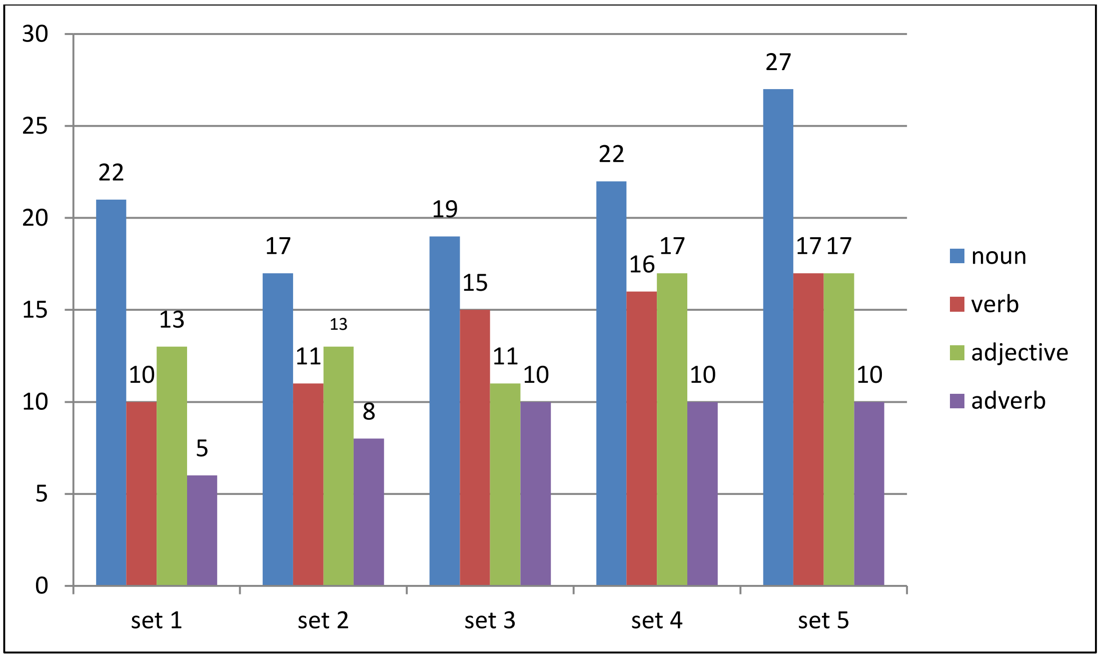


Figure 5.12: The proportion of words assigned to VSAT States 6+5 by set and class.

5.5.4 Notes on *individual* learning outcomes

Word gains of individual children varied substantially across sets. From Set 5, for example, participant 2 gained all four target words, ranking equal first (with several others), only to rank second to last from reading Set 4. Participant 7 emerged as the highest Set 4 scorer, along with participant 5, claiming four *known* words and yet placed among those gaining no words from having read Set 3. Within a set, the differences in *known* word totals could prove similarly substantial. For Set 1, participant 15 scored zero, while participants 5, 14 and 28 succeeded in learning 3 words each. From

Set 3, student 3 learned four words while student 6 gained just one. Although, in general, children gained additional words under the 6+5 test of *knowing* (the maximum gain for any child was 2), not all improved upon their State 6 based scores. Ten children failed to gain a higher sum of words from Set 1, for example, despite State 5 words now occupying the *known* category, with one child (participant 15) maintaining a zero score. From Set 2, nine children failed to gain additional words, two of whom retained a zero. Set 3 data reveals 7 non-gainers (three having also learned no words to the State 6 standard), and Set 4, 2 non-gainers, one of whom had failed to learn any words to the standard of VSAT State 6. Data for Set 5 stands in marked contrast, revealing that every child, other than participant 13, *knew* more words than they did under the stricter State 6 based test of *knowing*. Indeed, from this set alone, all 28 participants could claim to have learnt one or more words from their reading experience.

The across-sets totals of *known* words showed similar score variation. Participant 28 *knew* just four words to the State 6 criterion having completed every set of the five, though only 11 when *known* includes words in States 6+5. For participant 12, the *known* word total rises from 3 (State 6) to 12 and, for participant 5, from 5 to 17. Other notable gainers include participants 4 and 9 who successfully increased their *known* totals by 3 and 6 respectively. Cases in which a participant gained either more, or an equal number *known* words from reading the numerically higher designated set of the 5 remained few, with just four children displaying this outcome (participants 5, 10, 24 and 25). As under the State 6 notion of *known* word, no child's learning exhibited the reverse pattern i.e., a lesser or equal sum of *known* target words from reading higher numerically designated sets than lower.

5.6 *Known* words as those in VSAT States 6+5+4

5.6.1 States 6+5+4 raw data and statistical analyses

Table 5.3 and Figure 5.13 depict learning gains where a *known* word includes any occupying VSAT States 6, 5 or 4. A *known* word under this, the most expansive definition the study employs, satisfies one or more of the following tests:

1. The participant can use the word in a syntactically and semantically well-formed sentence, and supply a native like synonym.
2. The participant is able to provide a synonym for the target word (but not embed this in a clause), or
3. The participant believes (a.) s/he indeed *knows*, the target word to the required standard, and (b.) the administrator objectively considers this belief well-founded (see Section 3.5

A record of individual learning outcomes appears in Appendix 2c.

Set	Noun	Verb	Adj.	Adv.	average number of words learned by each participant	Percentage Learned of total (112)
1	23	17	16	13	(av=2.46); sd 0.92	62
2	21	19	21	11	(av=2.57); sd 0.74	64
3	24	22	19	16	(av=2.89); sd 0.92	72
4	24	23	19	14	(av=2.86); sd 0.71	71
5	28	25	22	17	(av=3.29); sd 0.66	82
	120	106	97	71		

Table 5.3: Known words (States 6+5+4) by set and lexical class (av=average; sd=standard deviation).

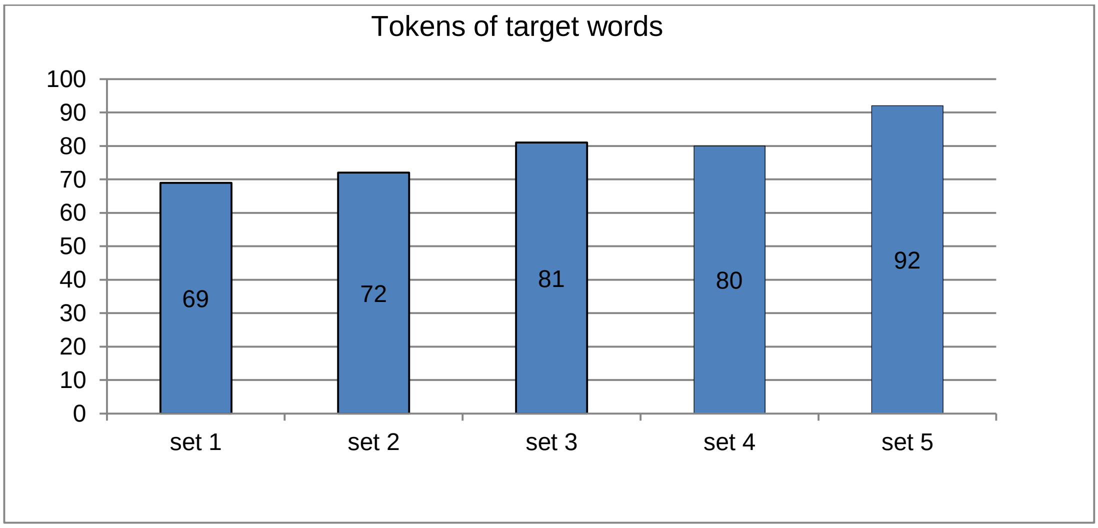


Figure 5.13: Totals of target words by set, States 6+5+4.

General comments

The highest sum of *known* words emerged from children having read Set 5 (92), followed by Set 3 (81), then Set 4 (80), Set 2 (72), and then Set 1 (69). Word gains over and above those recorded under the 6+5 test of *knowing* often proved substantial. For Set 1, children *knew* 18 additional words, an increase amounting to 35%. From Set 2, 23 additional words, representing an increase of 47%; from Set 3, 26 additional words, a 47% gain; from Set 4, 14 additional words, an increase of 23%; and from Set 5, 21 extra words, or 30% more than those occupying States 6+5 alone. Nouns once again proved the generally most learnable lexical class, as they had under both the State 6 and States 6+5 *known* word tests, children gaining more words of this type than of any other from all sets apart from Set 2; for the latter, *known* noun and adjective totals emerged as equal at 21 each. Adverbs again proved the least learnable class, as they had under the two alternative tests of *knowing* (see Sections 5.5 and 5.6, above). The totals of *known* verbs exceeded those for adjectives from reading Sets 1, 3, 4 and 5, while adjectives outnumbered verbs from reading Set 2. Participants' mean word gains ranged from a high of 3.29 per child, from Set 5, to a low of just 2.46 words from Set 1 (a difference of 0.83 words). The proportion of *known* words the participant group gained from any one set ranged from a low of 62% (from Set 1) to a high of 82% (from Set 5).

5.6.2 Research Question 1 (States 6+5+4): *How substantial, statistically, are differences in the number of known words from reading any one set as opposed to another?*

Findings:

The vote-count/sign test returned a p-value of 0.021 providing strong evidence for a 'pervasive' spacing effect contribution to differences in learned word sums (i.e. difference in medians) A Friedman's test (df=4, chi-square=19.268, p=0.001) and follow-up binomial (two-tailed) sign tests (Section 4.21) indicated significant 'pairwise' differences in *known* word totals associated with reading (a.) Sets 5 and 1 (p=0.004), and (b.) Sets 5 and 2 (p=0.000). With zeros shared between the N+ and N- categories, significant differences remained in each of these two cases: (Sets 5 and 1, p= 0.01; Sets 5 and 2, p= 0.00). No other significant differences were observed at either alpha =0.05 or 0.10.

Conclusion:

Findings from the vote-count/sign test provide 'strong' grounds for null hypothesis (i.e. no spacing effect; p=0.021) rejection, with the probability of observing the respective N+ and N- sums assuming the hypothesis correct amounting to fractionally over 1 in 50. The result of the 'pairwise' test procedure revealed two combinations of sets (5 and 1, and 5 and 2, of the 10 possible pairings) from which participants gained significantly more words from the relatively distributed learning condition. For both the 'general' and 'pairwise' tests, statistically significant p-values proved consistent with a spacing effect contribution to differences in children's learning outcomes (Section 4.21.1).

5.6.3 Research Question 2 (States 6+5+4): *How substantial, statistically, are differences in the sums of nouns, verbs, adjectives or adverbs from reading any one set as opposed to another?*

Statistical finding:

The vote-count/sign test identified a statistically significant difference in the sum of N+, N- cases for a single lexical class: verbs (p=0.002). For the remaining word classes the same test supplied the following: nouns, p=0.109; adjectives, p=0.344; adverbs, p=0.109. Cochran's Q tests for 'pairwise' differences in *known* sum totals failed to indicate significant variation in the relative sums for any of the four classes of interest (Figures 5.14–5.17) returning p-values ranging from p=0.109 (verbs) to p=0.551 (adverbs). Further McNemar tests prompted by Q to df ratios (Section 4.2), however, revealed a significant difference among paired proportions in three cases: (1.) in nouns from reading Sets 2 and 5 (p=0.016); and (2.), verbs from reading Sets 1 and 5 (p=0.039); and (3.) verbs from reading Sets 2 and 5 (p=0.031).

Conclusion:

The 'general' analysis, drawing upon findings from the vote-count/sign test procedure, allows null hypothesis rejection (no spacing effect) for a single word class, verbs, implying a positive association between longer target word presentation time and more substantial learning i.e. *known* word totals. Results from the McNemar 'pairwise' comparisons, however, indicated the participant group learned proportionally more nouns and/or verbs from the set providing the relatively spaced learning condition in three cases: (1.) The pair of sets 2 and 5 (nouns), (2.) the pair 1 and 5 (verbs), and (3.) the pair of sets 2 and 5 (verbs). The null hypothesis (Section 4.22.1) of a zero spacing effect is rejected, leaving differences in target word presentation time a plausible cause of the additional word gains.

N	28
Cochran's Q	7.222
Df	4
Asymp. Sig	0.125

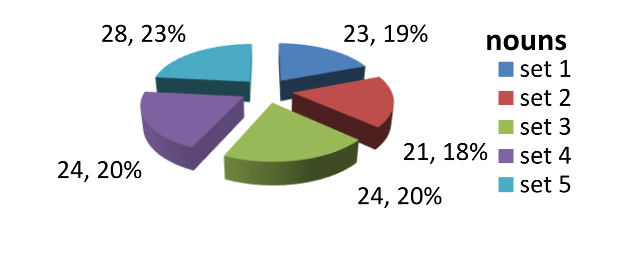


Figure 5.14: Percentage of nouns assigned to States 6+5+4, by set of texts.

N	28
Cochran's Q	7.556
Df	4
Asymp. Sig	0.109

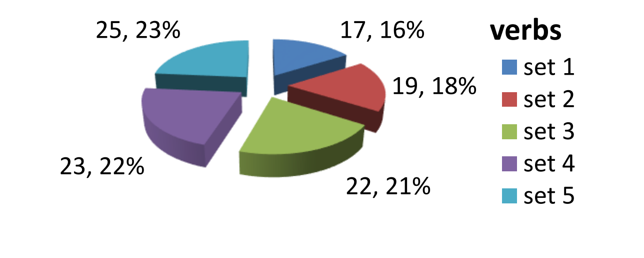


Figure 5.15: Percentage of verbs assigned to States 6+5+4, by set of texts.

N	28
Cochran's Q	3.533
Df	4
Asymp. Sig	0.473

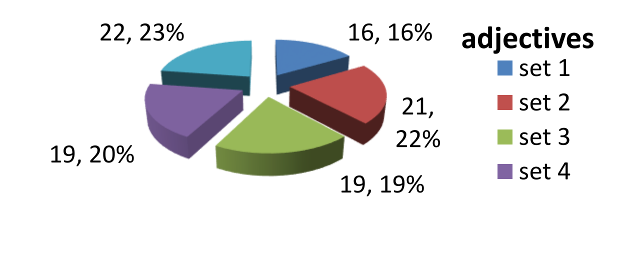


Figure 5.16: Percentage of adjectives assigned to States 6+5+4, by set of texts.

N	28
Cochran's Q	3.04
Df	4
Asymp. Sig	0.551

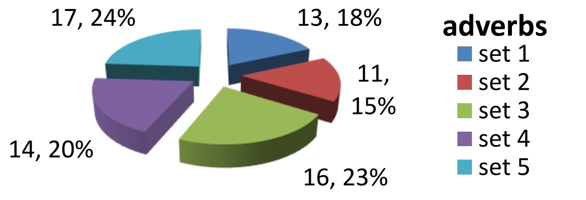


Figure 5.17: Percentage of adverbs assigned to States 6+5+4, by set of texts.

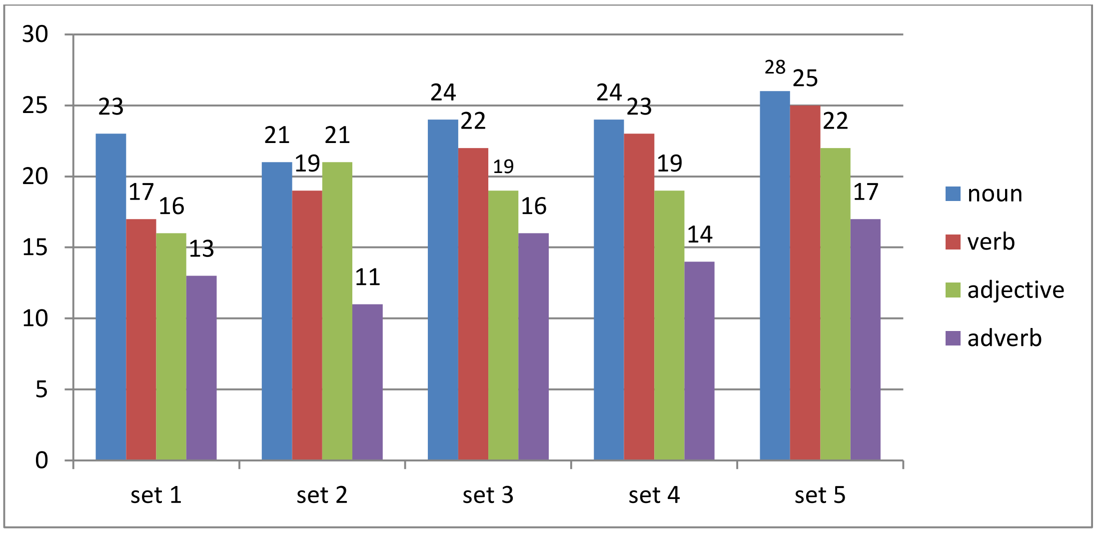


Figure 5.18: The proportion of tokens assigned to VSAT States 6+5+4 by set and lexical class.

5.6.4 Notes on individual learning outcomes

The variability in learning outcomes proved 'substantial.' From Set 5, 11 learners (from the 28) gained all 4 target words, while three gained just two (no one gained one or zero). For Set 4, *known* word totals ranged from 1 (a single participant) to 4 (4 participants), all children having gained at least one word from their reading experience. Seven children learned all 4 words from Set 3, while one (participant 27) failed to learn any of the potential four. Across sets, outcomes proved similarly variable. Participant 27 *knew* zero words from Set 3, ranking last among the 28 children, and yet 4 words from Set 5, ranking joint 1st. Compared to the learning observed under the State 6 based test of *knowing,* children's gains proved generally 'large.' Participant 27, for example, *knew* just 2 words to the State 6 standard having read all 5 sets of experimental texts, yet 12 if *known* words include those in States 6+5+4, an increase of 500%. For student 12 the *known* word sum increased from 3 (State 6) to 17 (States 6+5+4), a rise of 467%. None of the children learned no words under the 6+5+4 test of *knowing* from reading any of the Sets 1, 2, 4 and 5, and just 1 (participant 27) from having read Set 3. Eight participants consistently gained either more or an equal sum of words from the higher numbered set of a possible pair (e.g. Set 3 as opposed to Set 2) while no child displayed the reverse pattern of invariably fewer or equal *known* words from the set offering the more massed target word presentations. The total of children successfully gaining all four words embedded in a particular Set amounted to 4 from having read Set 1; 2 from reading Set 2; 7 from Set 3; 4 from Set 4; and 11 from Set 5.

5.7 Summary, Part 1

Employing two types of analyses, 'general,' and 'pairwise,' the study demonstrates that the time intervals between readers' reencounters with target non-words in sets of texts designed to the same specifications could result in statistically significant differences in (a.) the sums of *known* words children gained from their reading experiences (the concern of Research Question 1), and (b.) the number of nouns and/or verbs, from reading one set as opposed to another (the concern of Research Question 2). The study quantified the vocabulary gains from reading each set in terms of *known* word sums, average gains, and percentages, as well as reporting specific differences in learning outcomes from reading the sets of a possible pairing. Pairs of sets from which children collectively gained statistically significantly more words overall (undifferentiated by class) or of a particular lexical category from reading one set of the two, proved relatively few. However, where *pairwise* and/or *general* testing allowed for null hypothesis rejection, the evidence often came across as compelling, yielding p-values in some instances falling below 0.01. Despite occasional indications of a spaced learning advantage, notes on individual children's learning caution that conclusions applicable to the participant group only poorly predict any single child's likelihood of word gain regarding (a.) the total words s/he might learn (i.e. undifferentiated by class), or (b.) total gains of a particular lexical class from the set from which his or her learning proved most substantial. The study notes several instances where a child would learn fewer words in 'aggregate,' or of a class, from sets from which the participant group gained significantly more. Despite signs that spaced learning reasonably accounts for participants ‘picking up’ additional words from sets presenting more distributed learning opportunities, few children displayed a pattern of consistently more word gains (or equal number of learned words) from reading the numerically higher ordered set of a pairing. No child, however, displayed the opposite pattern (suggestive of a massed learning advantage) of consistently less or equal gains from the numerically higher set.

Part 2 (further statistical analysis)

5.8 Interpreting statistical differences in learning outcomes

Part 1 identified cases of statistical differences ($p<0.05$) in word gains associated with children reading alternative sets of texts, each set so designed that it differed from the others according to the time over which a reader encountered unique (to that set) embedded target words. In revealing such cases the study affirmed instances of a spacing effect contribution to learning within the context of the participant population and reading circumstances under which the present investigation took place. The current discussion (Part 2) moves on to examine each of these several cases (pairings of sets associated with null hypothesis rejection) in rather more detail. The discussion addresses two issues: (1.) What do significant ($p<0.05$) differences amount to in quantitative terms? and (2.) What is the distribution of words among the particular VSAT States which define the particular sense of *knowing* of interest?

Research Question 1

5.8.1 *Known* words as those that occupy VSAT State 6 (Research Question 1)

Under this, the most stringent test of *knowing*, differences in word gains ranged from a high of 7 (from reading Sets 5 and 3) to a low of 0 (Sets 1 and 4). Friedman tests (Section 5.4) for 'pairwise' comparisons identified no cases in which the participant group gained more words from any set of a possible pairing leaving the null hypothesis of no apparent spaced learning advantage firmly intact. A second round of testing employing the vote-count/sign test, likewise returned a non-significant value (p=0.753), corroborating the pairwise test findings. Both the 'general' and 'pairwise' tests point towards a common finding: Children, collectively, did not gain more vocabulary to the productive competence VSAT State 6 captures from longer duration time intervals between target word presentations.

Conclusion 1: With *known* words defined as those a participant could supply in a syntactically and semantically well-formed clause, children (as a group) did not gain statistically more target words whether they encountered those words under more or less distributed learning conditions.

Conclusion 2: Of the ten possible pairs of sets, given the five sets the study employs, 3 pairs are associated with a higher sum of target words in State 6 from the set offering the more massed learning experience, and 6 pairs with numerically more words from the set offering the relatively spaced learning condition. The difference (N-=3; N+=6) is not statistically significant.

From the association of VSAT States with particular lexical competencies:

Conclusion 3: More, or less, spaced encounters with target words during RR sessions did not have obvious benefits, failing to yield significantly (p<0.05) more *known* words of the type children *might apply productively* to generate semantically and grammatically well-formed clauses.

A note on statistical validity

Conclusions remain tentative given the well-documented sign test insensitivity for detecting a median divergence from zero should relatively few paired samples comprise the data source. To categorically discount a possible spaced learning advantage had a larger body of children participated in the research would be an error, as would rejecting the a positive contribution of spaced learning had the investigation employed sets of texts presenting target words over more than five days. A hint that the data may, contrary to test findings, indeed conceal an operative spacing effect comes from the general texture, or profile, of the data as appears in Figure 5.1. The 'visually evident' transition (see

Figure 5.1) from the participant group gaining rather more words from the massed learning associated with Sets 1 and 2, compared to words learned from Set 3, stands out –so, too, the unmistakable 'suggestion' of a spaced learning advantage from children having gained rather more words from Set 5 than from Set 4 and, then again, more from Set 4 than Set 3. Set 3 appears to stand as something of a figurative 'divisor,' or fence, before which (Sets 1 and 2) children gained additional words from ever more massed learning (i.e. from stepped reductions in word presentation time), and beyond which (Sets 4 and 5) we have the faint indication of a spaced learning advantage with larger word gains associated with numerically higher ordered sets. Sign-tests, a non-parametric 'standard' for comparing medians, have no sensitivity to such data patterns given the bare sums (signs) upon which they 'operate' (e.g. simple totals of N+ and N- values). That is, they fail to acknowledge the magnitude of differences between data values,[87] The possibility of a spaced (or massed) learning effect impacting upon word learning, from this perspective, remains, the suggestion being that more distributed or massed presentations account for disparities in target word uptake depending upon whether the sets one chooses for comparisons lie to the 'right,' or to the 'left,' of the metaphorical Set 3 'divisor.' That children (collectively) gained more words in total, and by progressively larger margins, with each 24-hour presentation time extension beyond 3 days (i.e. the 72 hours, Set 3 provided) is the very outcome that an operative spacing effect may account for. Prior to the three day (72 hour) mark, the suggestion becomes rather more of a massed learning benefit, the participant group having gained additional words from each 24-hour presentation time reduction. The hypothesis of a spaced learning advantage, initially absent, that subsequently begins to express itself once children encounter target words over more than three days supplies a plausible, albeit tentative, explanation of the wave-like 'fall (massed learning advantage) and rise' (spaced learning advantage) profile apparent from 'reading' from left to right across the columns of the Figure 5.1 display.

[87] A sign test only looks at the sign (+ or -) of the 'difference' between a pair of values –in other words, whether values are higher or lower than others, without regard to by how much they may differ. The test is insensitive to the apparent 'implications' of a directional trend we see reading across the top row of Figure 5.1.

5.8.2 *Known* words as those that occupy VSAT States 6+5 (Research Question 1)

Extending the *known* word class to include those in VSAT State 5 in the *known* word category, resulted in both 'pairwise' tests and the vote-count/sign test (p=0.021; two-tailed) supplying compelling evidence for word gains from distributed target word presentations. A 'pairwise' sign test (Part 1) indicated a significant difference in *known* totals in four cases after apportioning tied scores (Section 5.5.2):

1. case (**a**$_{, 6+5, 5\&1}$): from reading Sets 5 and 1, (p=0.01)
2. case (**b**$_{, 6+5, 5\&2}$): from reading Sets 5 and 2, (p=0.01)
3. case (**c**$_{, 6+5, 5\&3}$): from reading Sets 5 and 3, (p=0.01)
   and
4. case (**d**$_{, 6+5, 4\&2}$): from reading Sets 4 and 2, (p=0.04)

5.8.3 The source of significant differences

The source of the statistical differences becomes reasonably clear from the distribution of additional *known* words among VSAT States 6 and 5, these states defining a more expansive *known* word class (Figure 5.19) than occupancy of State 6 alone. A paired sign test (prior to apportioning zeros) for 'pairwise' differences in State 5 word sum totals returned highly significant, and identical, 2-tailed p-values of 0.007 for cases (**a**$_{, 6+5, 5\&1}$) and (**b**$_{, 6+5, 5\&2}$) alike. After distributing zero scores between the N+ and N- 'columns' the values rise to p=0.038 and p=0.014 respectively, though still falling comfortably below the conventional alpha (0.05). The findings allow a confident null hypothesis rejection and a strong assertion that target word presentation time indeed explains the larger State 5 sums from the Set 5 reading experience. A follow up vote-count/sign test for a spaced learning advantage taking as input the sums of pairs of sets from which children gained more (N+ =9) or less (N- =1) target words to the State 5 *known* criterion yielded a p-value of 0.021, again highly indicative of a spaced learning contribution to differences in *known* (State 5) word totals; the probability of the sums of N+ N- having arisen by chance were the null hypothesis correct now only marginally exceeds 1 in 50.

For case (**c**, $_{6+5,\ 5\&3}$), results from sign testing to identify ‘pairwise’ differences proved ‘surprising,’ revealing respective State 5 sums from Sets 5 and 3 missing statistical significance comfortably (p>0.21, before allocating zero scores) despite children gaining rather more words overall (i.e. the sum of those in States 6 and 5) from Set 5 (p=0.01), and binomial sign test evidence affirming a spaced learning explanation for differences in State 5 totals in both cases (**a**, $_{6+5,5\&1}$; p=0.007), and (**b**, $_{6+5,5\&2}$; p=0.007). The source of the anomalous case (**c**, $_{6+5,\ 5\&3}$) finding remains unclear; the most probable cause –albeit tentatively proposed–best ascribed to one, or the other, of two factors that explain non-significant p-values generally, namely: (1.) An effect truly absent (that is, children did not learn more from distributed target word encounters; this amounts to the null hypothesis position), or (2.) That despite a spacing effect contribution to differences in *known* State 5 word sums, binomial sign test procedures display insufficient power to detect any such effect given the limited input data (i.e. the respective totals of State 5 words from reading Sets 5 and 3).[88] It remains uncertain which of these options amounts to the more plausible explanation. The ‘low power’ hypothesis garners some support from the unambiguous evidence spaced learning indeed accounts for the extra State 5 words participants gained from reading both Sets 5 and 1, and Sets 5 and 2,[89] together with the vote-count/sign test indication of a spaced learning advantage in regard to words children *knew* to the State 5 standard (as noted, p=0.021, see p.167). The findings raise an intriguing issue of data interperetation and reconciliation: Just how to account for the participant group having learned significantly more State 5 words from Set 5 over those from Set 3, and yet failing to learn significantly more from reading Set 5 as opposed to Set 2, despite Sets 3 and 2 providing only a 24 hour difference in target word presentation time. Unless we suppose an abrupt spacing effect curtailment –that is, an effect that does not arise should a set offer merely 24 hours additional learning time– it might seem reasonable to presume a spaced learning advantage accounts for the extra State 5 words in case (**c**, $_{6+5,\ 5\&3}$), the high p-value

[88] For example, from tossing a coin 3 times and getting two heads and one tail is not a significant difference in results p=1.00); sign tests would, however, return a significant finding from 150 coin tosses which supplied a 100 heads and 50 tails (p=0.001), albeit the ration of heads to tails is the same, 2:1.
[89] In other words, that children indeed gained significantly more words from set 5 in these cases.

(p<0.21; alpha =0.05) notwithstanding. The percentage contribution of additional State 5 words to the total words gained from the set offering the more spaced learning (Set 5) could prove substantial. In case (**a**, $_{6+5,\ 5\&1}$), the extra State 5 sums (16) make up 80% of that addition. In case (**b**, $_{6+5,\ 5\&2}$), this contribution is an impressive 72.7%.

That children learned 6 extra State 6 words from Set 5 compared to those gained from reading Set 3 lies consistent with a spacing effect securing the larger State 6 sum, albeit arguably is not so impressive a word gain as to preclude mere coincidence as a plausible cause of the difference in totals (Section 5.8.1). Whether the modest addition to State 6 words from having read Set 5 truly stems from spaced learning in case (**c**, $_{6+5,\ 5\&3}$) remains unclear. The results from both 'pairwise' sign tests and the 'general' test procedure supplied no grounds for null hypothesis (i.e. no effect) rejection (Section 5.8.1), yet fall short of conclusive given (a.) the sign test bias towards null hypothesis preservation when few observations serve as input, and (b.) tentative interpretations we can draw from the prominent 'U' shaped profile readily obervable in Figure 5.1 (p.148) to which binomial sign test findings prove insensitive. From the wave-like 'pattern' of ever larger additions to State 6 sums accompanying each 24-hour extension to presentation time exceeding 72 hours, the signficant difference in *known* word sums in case (**c**, $_{6+5,\ 5\&3}$) plausibly arises, in part at least, from a general increase in words familiar to both the State 5 and State 6 standard (though more so sums occupying the former State) from the additional presentation time Set 5 afforded. This supposes (and controversially) a very real spacing effect but one falling below the threshold of detection given inherent sign test limitations and the limited data available serving as input

<u>Conclusion 4</u>: With *known* words defined as those that occupy VSAT States 6+5, children could gain statistically (p<0.05) more from either: (a) a general increase in words familiar to the standard of those in States 6 and 5 – see e.g. case (c,$_{6+5,\ 5\&3}$) or (b) from a disproportionate excess in the sums of State 5 words from the set affording the more distributed learning condition; the examples cases are (a, $_{6+5,\ 5\&1}$) and (b, $_{6+5,\ 5\&2}$).

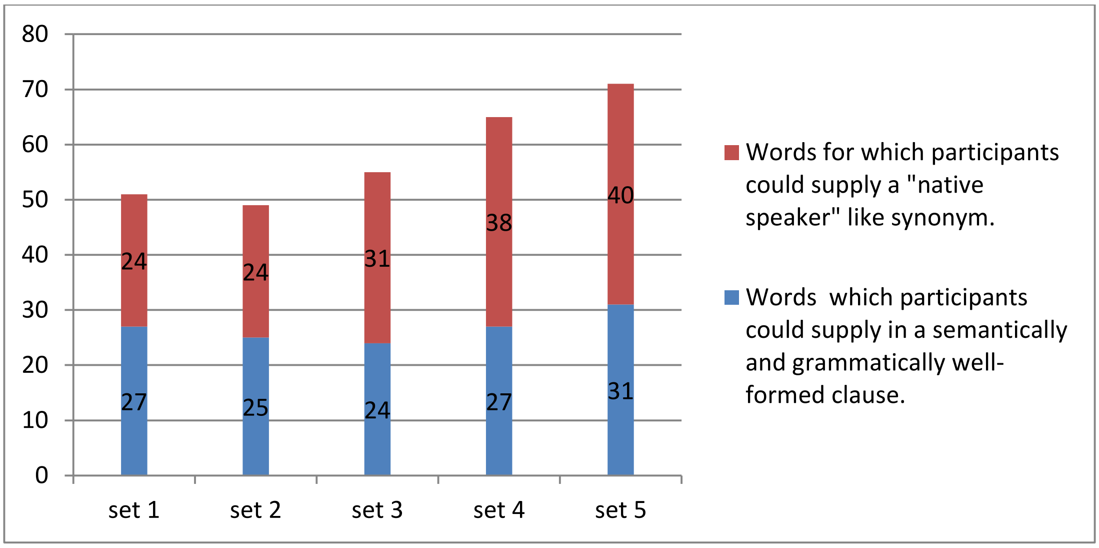


Figure 5.19: Proportion of words in States 6 and 5 cases (**a,** 6+5, 5&1), (**b**, 6+5, 5&2), (**c**, 6+5, 5&3) and (**d**, 6+5, 4&2).

Case (**d**, 6+5, 4&2) stands as the only pairing of sets from which children collectively gained the significantly larger *known* word sum (p=0.04) from other than reading Set 5 –a total of 16 extra words in this instance from their Set 4 reading experience over those learned from reading Set 2. A sign test ('pairwise') to identify a spacing effect contribution to differences in the respective VSAT State 5 sums returned a significant value (p= 0.031) before zero apportioning, and a non-significant p=0.051, thereafter; the latter figure misses significance by only the narrowest of margins (0.001), only barely ruling out 'kinship' with cases (**a**, 6+5, 5&1) and (**b**, 6+5,5&2) from which readers gained statistically more State 5 target words from relatively spaced word encounters. Case (**d**, 6+5, 4&2) differs from the latter two cases in that spacing appears less likely explains disparities in *known* State 6 sums given that the sums in this state only 'noticeably' increase with 24-hour increments to presentation time beyond the three day mark i.e. 72 hours (see Figure 5.1 and Section 5.8.1). If, as previously cautiously suggested, Set 3 indeed marks a divisor beyond which children gain progressively more words with each 24-hour presentation time addition, then extra words familiar to the State 6 standard from Set 4 in case (**d**, 6+5, 4&2) should prove numerically lower than from children having read Set 5, Set 4 offering the less optimal spaced learning opportunity; this is indeed borne out in the sums of words children *knew* to the State 6 standard, as noted (Section 5.4.1).

The pair of Sets 1 and 4 represents a noteworthy example of children having gained statistically more State 5 words (p=0.049) before apportioning zero scores equally between the N+ and N- observations, despite that Friedman and follow up standard sign tests failed to identify a significant difference in gross (undifferentiated by State) *known* (6+5) word totals (see Part 1). After zero apportioning the p-value rises to 0.19, comfortably missing statistical significance. Data from Sets 1 and 4 seem to suggest –albeit not quite affirming– that how many words the participant group gained to the State 5 standard could differ significantly (p<0.05) despite no such difference in gross *known* word sums (the aggregate of those in States 6 *and* 5) from reading one set or another (p=0.09). The observation would have escaped notice had the study not subdivided each notion of *knowing* into its own discrete VSAT based lexical competencies (i.e. combinations of VSAT States). The finding prompts an extension to conclusion 4:

Conclusion 5: Should the sum of *known* words overall (the totals of words occupying VSAT States 6 and 5) not differ signficantly from reading one set of a pair, children could nevertheless gain statistically more target words to the VSAT State 5 standard from a specific set of a pair offering the more distributed learning opportunity.

More generally:

Conclusion 6: The effects of spacing disproportionately impact upon the sums of words *known* to certain lexical competencies more so than others.

5.8.4 Differences in target word presentation time on learning outcomes

Cases (**a**$_{,\ 6+5,5\&1}$), (**b**$_{,\ 6+5,5\&2}$), (**c**$_{,\ 6+5,5\&3}$) and (**d**$_{,\ 6+5,\ 4\&2}$) help shed light on spacing effect contributions to word gains through the hints as to how substantial discrepancies in presentation time need be before children gain significantly more words from a set of a pair. For gross *known* word sums (the sum of those in States 6 and 5), and limiting discussion to State 5, the minimum addition to presentation time sufficing before 'pairwise' tests identify significantly (p<0.05) more learning amounts to 48 hours. That this was not always sufficient in itself to ensure significant differences, however, is evident from children's failure (as a group) to gain statistically more words from Set 3 as opposed to Set 1 despite the same 48-hour presentation time disparity. Whether, for any particular pair of sets, children collectively gained significantly more words from one or the other would, it appears, depend not just upon presentation time but rather the interplay of two factors: (1.) *difference* in presentation time (whether a set provides 24, 48, 72 or 96 hours of extra learning time over that of another), and (2.) *actual* presentation each set of a pair required for completion –i.e. the number of days called for to read a set's complement of texts in their entirely (i.e. a number of between 1 and 5). The *difference* in presentation time from reading both Sets 4 and 2 ($\mathbf{d}_{,\,6+5,\,4\&2}$) and Sets 3 and 1, for example, amounts to the same i.e. exactly 48 hours. The *actual* presentation time is dissimilar –namely, 2 days and 4 days in the case of Sets 2 and 4, as opposed to 1 day and 3 days for the particular pair of Sets 1 and 3.

Presentation time extensions of 24 hours did not result in significant differences in *known* word totals (the sum of words in States 6+5) from reading any set of a pair, whatever the actual time over which presentation occurred. The results from pairwise binomial sign tests (even after allocating zeros) proved conclusive, returning p-values ranging from p=0.286 to p=1.000 for all pairs of sets to which the tests were applied.[90] These findings strongly buttress the null hypothesis position (no spaced learning advantage) albeit subject to the caveat of a possibly contrary result had the study supplied more sets for testing (perhaps sets of texts presenting target word encounters over 6,7, or 8 days) and a larger participant pool. The more general issue arising from the 24-hour data concerns the question of 'boundary' or 'gradation:' i.e. whether we have a minimum addition to presentation time (a figure somewhat less than 48 hours but exceeding 24) below which spacing proves inoperative (a boundary of sorts), as opposed to an effect falling below detection yet nonetheless reasonably explaining differences in learning outcomes. The notion of boundaries defining time points beyond which language acquisition fails was noted in Chapter 2 (Section 2.5.2) in reference to critical ages after which gains of particular linguistic competencies appear improbable or even impossible. Specific time points in months or years after birth at which certain capacities or aptitudes begin to emerge, or fail to develop further, have received much attention in the child learning literature. The onset of abstract

[90] These pairs are, namely, sets 1 and 2; sets 2 and 3; sets 3 and 4; and sets 4 and 5.

thinking, to take one example, only usually occurs when children reach around 12 years of age. In works discussing the spacing effect, the notion of 'boundaries' attracted some attention in Janiszewski et al. (2003) and more recently in Rohrer (2009). Whether children in the present study might indeed have gained statistically significantly ($p<0.05$) more vocabulary from additions to target word presentation time of just 24 hours remains indeterminable, however, given the small-scale nature of the investigation –just 28 EAL students. Findings from the present research that might bear upon the boundary issue remain equivocal, the data neither ruling out nor affirming 48 hours as a minimum extension to presentation time below which significant learning differences invariably fail to arise. Nor can one reasonably comment upon the pedagogical value of word gains should 24 hours indeed suffice to ensure a learning advantage –i.e. the relative dispersion of additional words among States 6 or 5. The small-scale nature of the present study yields insufficient data for analysis. The issue of boundary or gradation in regard to the time intervals between word reencounters and word gains in the context of recreational reading calls for further research and, especially so, investigations able to recruit somewhat more children than proved possible for this dissertation project. Ideally, such future studies might employ the same vote-count/sign test procedure, and pairwise testing as employed in the present research.

Despite several examples of the participant group gaining significantly ($p<0.05$) rather more words (whether of State 5, or the sum of those in States 6 and 5 combined) from sets presenting target vocabulary over 48, 72 or 96 additional hours, the small-scale nature of the research precludes an inference of more impressive learning had word presentation time exceeded the maximum (5 day) period examined. Nor does the study, and for the same reasons, make claims as to the maximum time interval between the 12 encounters with the same novel word beyond which spaced learning might yield diminishing returns to, ultimately, become less learning conducive than massed. This 'switch over' to more learning from massed –as opposed to spaced– target word presentations would most plausibly arise once intervals between meeting the same word become sufficiently large that the reader can no longer recall meeting the word on a prior occasion (Pimsleur, 1967). The number of days that must elapse, beyond the apparently optimal 5, such that further 24-hour time additions would induce ever smaller spaced learning gains would likely vary from child to child, albeit for populations general 'patterns' should prove discernable, just as they do for the participant group in the present study. The question remains as to whether a child, or group of children, would gain more words from 12 encounters distributed over longer periods e.g. 6, 7 or 8 days, than form five. The data on individual learning outcomes (Sections 5.4.4, 5.5.4 and 5.6.4) raises this possibility (see Chapter 6, Section 6.3.2) –at least for some children– as do the obvious differences in children's aptitudes for remembering and learning declarative knowledge generally (Paradis, 2009). In any typical classroom, you find students who more ably grasp and retain information than others, a detail often obscured when considering a participant group as a whole. What amounts to the optimum degree of spaced learning for one child may not be so for his or her peers.

5.8.5 Lexical competencies

To examine particular lexical competencies takes the discussion of spaced learning beyond the gross sums of words occupying a VSAT state, or states, to explore what gains amount to in terms of productive and/or receptive understandings –it shifts attention, that is, to the function or purpose to which a child can apply newly learned words as opposed to dwelling upon bare numerical differences in *known* word totals. What, then, does the study reveal of children's lexical competence gains?

The study, arguably, denies a relationship between presentation time intervals and children's capacity to supply target vocabulary in 'grammatically and semantically well-formed clauses' –the competence VSAT State 6 claims to capture. On this, the results of 'pairwise' and 'general' tests unambiguously agree, each supplying p-values markedly exceeding the $p<0.05$ that would have confidently allowed for null hypothesis rejection. Further research into the duration of target word presentation time on State 6 sums would, however, seem a worthy pursuit given both the undoubted pedagogical importance of the competence the state claims to capture, and the mild hint of a spacing effect noted in Section 5.8.1. If, however, as the test findings clearly suggest, State 6 sums indeed prove unresponsive to spacing this leaves the capacity 'to supply a native-speaker like synonym' (i.e.

the skill VSAT State 5 designates) as the sole lexical skill, of the two examined, apparently reactive to manipulating time intervals between the same target word's textual occurrences. While this skill (obviously) falls well short of the knowledge that suffices to supply words in syntactically and semantically well-formed clauses, it would usefully allow simple responses to direct questions, enable expression of basic needs, and make a positive contribution towards a child's comprehension of oral or written communication. In school settings, *knowing* a word to the State 5 standard might helpfully assist in performing such typical tasks as completing multiple-choice (vocabulary) tests, understanding the gist of conversational exchanges, following lessons, and making sense of written scripts, including textbooks. Despite room for teacher disagreement as to, for example, the significance of alternative lexical competencies (See e.g. Section 1.4), and the relative importance of receptive and productive skills, the advantage of Set 5 type texts for promoting child vocabulary development stands out, nevertheless. Only from Set 5 did children collectively gain statistically significantly more words than those from having read an alternative set.[91] Furthermore, and irrespective of whether we define *known* words solely in terms of State 6 placement, State 6+5, or occupancy of States 6+5+4, children gained more words from Set five than from reading any of the remaining four (See Appendix 2a, 2b, and 2c).

<u>Conclusion 7</u>: Of the two lexical competencies which VSAT States 6 and 5 attempt to capture, only the capacity to supply a native-like synonym (indicated by word occupancy in State 5) proves unambiguously responsive to whether learning occurred under a more, or less, spaced condition.

<u>Conclusion 8</u>: Manipulating the time intervals between encounters with novel words in any one set of texts is unlikely to result in significant differences in the stock of words readers can supply in syntactically and semantically well-formed clauses.

5.8.6 Known words (6+5), and measures of gain; Research Question 1

Figure 5.20 reports differences in the sums, averages and percentages of *known* words from reading the set of each possible pair associated with statistically significantly more learning. From

[91] The issue of how substantial are word gains over meaningful time periods, e.g. one academic year is the subject of Part 3 of the present chapter.

reading Set 5, for example, children learned 20 words more than from reading Set 1, amounting to a 39.2% increase in the participant group's *known* total. Children gained 16 more words from reading Set 4 compared to Set 2, corresponding to an average 0.57 extra words per child (out of a maximum 4 possible words potentially gainable), and 16 extra from reading Set 5 over Set 3. The maximum gain reasonably attributed to differences in target word presentation time (as derived from McNemar tests) amounted to 0.79 additional words (out of 4) per child, this achieved from having read Set 5 as opposed to Set 2; the minimum stood at 0.57 extra words, this from reading Set 4 as opposed to Set 2.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					Differences in sums: 20 Differences in the averages of words gained: 0.72 Percentage increase in words gained: 39.2
Set 2				Differences in sums: 16 Differences in the averages of words gained: 0.57 Percentage increase in words gained: 32.8	Differences in sums: 22 Differences in the averages of words gained: 0.79 Percentage increase in words gained: 44.8
Set 3					Differences in sums: 16 Differences in the averages of words gained: 0.58 Percentage increase in words gained: 29

Figure 5.20: Differences in the averages, sums and percentages of *known* (6+5) target words.

5.9 *Known* words as those that occupy VSAT States 6+5+4 (Research Question 1)

Under the least restrictive definition of *known* (Occupancy of VSAT States 6, 5 and 4), Friedman tests (and post hoc sign testing having apportioned zero scores) indicated a significant difference in *known* word sums in 2 cases:

1. case (**a**, $_{6+5+4,\ 5\&1}$): from reading Sets 5 and 1, (p= 0.01)

and

2. case (**b**, $_{6+5+4,\ 5\&2}$): from reading Sets 5 and 2, (p= 0.00)

5.9.1 The source of significant differences

The source of these differences becomes apparent from Figure 5.21 which depicts the totals of *known* words occupying each of the VSAT States 6, 5 and 4 for the constituent sets of the pairs (i.e. Sets 5 and 1, and Sets 5 and 2) of interest. The sums for States 5 and 6 occupancy received attention in Sections 5.8.1 and 5.8.3 during the discussion of significant differences in learned word totals arising under the 6+5 notion of word *knowing*. This leaves only the State 4 sums to have escaped attention thus far. What, then, do the 'general' and 'pairwise' tests reveal of words familiar to the VSAT State 4 standard? The results from the vote-count/sign test proved unambiguous, offering no support for a 'general' spaced learning advantage from the respective N+ N- values. Of the 10 possible pairs of sets, children gained additional State 4 words from that offering the more massed learning in 5, and more from the set providing the relatively distributed learning in the remainder (no ties). The sums of N+ and N- cases (i.e. 5 and 5) comfortably miss significance (a sign test returned a p-value of 1.000) leaving the null hypothesis intact (Section 4.21.1). The 'pairwise' test results proved no less definitive. For cases (**a**, $_{6+5+4,\ 5\&1}$) and (**b**, $_{6+5+4,\ 5\&2}$) a binomial sign test applied to the respective State 4 sums (zeros apportioned) returned non-significant p-values (p=1.000) for both Sets 5 and 1, and Sets 5 and 2.

The insensitivity of State 4 *known* word totals presentation time suggests the significant (p<0.05) differences observed in gross totals in both cases (**a**, $_{6+5+4,\ 5\&1}$) and (**b**, $_{6+5+4,\ 5\&2}$) stem from the same source as gains observed under the States 5+6 sense of *knowing*: namely, the additional words learned to the State 5 standard arising from the relatively spaced learning that the set provided, along with any extra State 6 words, assuming sums in the latter state indeed exhibit some sensitivity to presentation time manipulation (the issue remains controversial as we saw in Section 5.8.1). With words occupying State 5 comprising less of the *known* sum under the 6+5+4 definition of *known* given apparently 'spacing insensitive' State 4 words now falling within the *known* category one might suppose a lesser likelihood of significant differences (p<0.05) in the sums of words in States 6+5+4 from set-to-set comparisons than under the State 6+5 notion of knowing.[92] The somewhat anomalous

[92] The claim is that were the sums of words in state 4 a larger proportion of the totals of the respective *known* word totals

finding that 'pairwise' sign testing (after apportioning zeros) returned roughly comparable p-values for pairs of sets 5 and 1, and 5 and 2, under both the 6+5 and 6+5+4 senses of *knowing* likely stems from the small number of words occupying State 4 relative to those familiar to the 6+5 '*known word*' standard (see Section 5.8.2) in each of the Sets 5, 2, and 1. The comparatively minor sum, in this view, essentially prevented expression of the logical effect of including 'insensitive' State 4 words in the *known* class –i.e. a reduction in the number of instances in which a child would gain additional words from the set providing the more spaced learning opportunity.

Conclusion 9 With *known* words defined as those occupying States 6+5+4, the source of statistical difference in the totals of *known* word sums (should they arise at all) lay primarily in the increase in words occupying State 5 from reading the set of a pair offering the more distributed learning condition (Set 5).

Conclusion 10: In terms of words *known* to the particular standard of either VSAT State 6 or State 4, relatively distributed learning does not result in statistically more gains from the set affording the more distributed learning condition (Set 5).

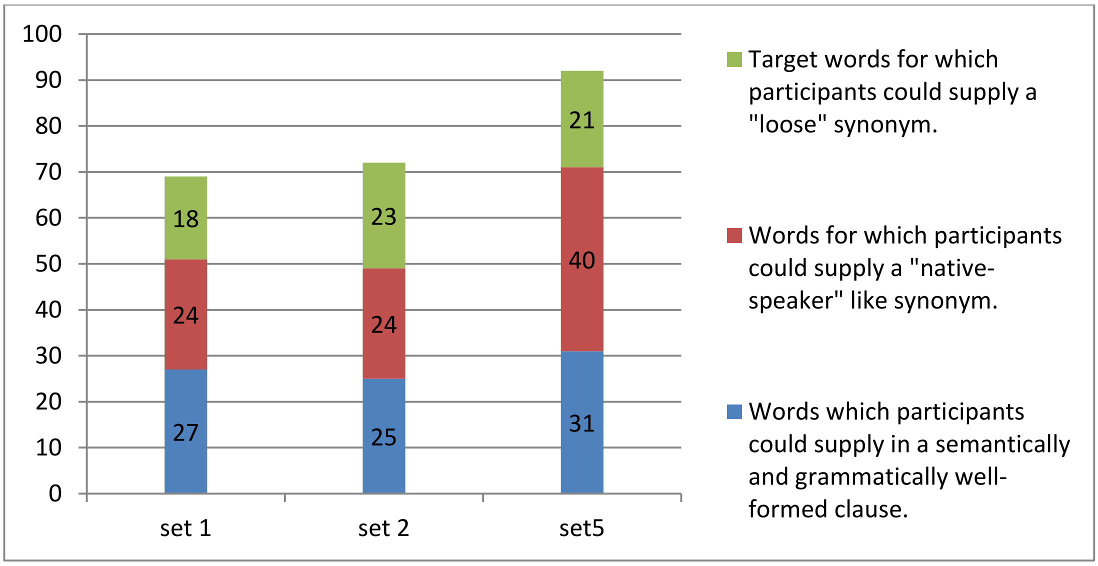


Figure 5.21: The proportion of words in States 6, 5 and 4; (**a**, 6+5+4, 5&1) and (**b**, 6+5+4, 5&2).

from each set of a pair, the p-values from sign tests on those totals would prove lower than were state 4 not included in the *known* sum (i.e. if the *known* sums included only words in States 6+5). This arises from the lesser number of cases in which children gained more words from spaced learning than under the state 6+5 definition of *knowing* compared to the 6+5+4 definition. To illustrate: Given two columns of matched figures, those in the second column all containing higher values than the first, randomly increasing the value of the figures in column 2 fails to raise the number of cases in which totals in the second exceed the matched figure in the first (the 'adding' merely increases the difference between them). Conversely, randomly adding figures to the first and second column may on occasion result in totals exceeding those of Column 2, depending upon by how much the column 1 figure increases.

5.9.2 Lexical competencies

Moving on to consider just what children *knew* of any additional words gained (as opposed to how many), two points stand out: First, we have clear evidence (Figure 5.21) that the participant group failed to learn statistically more words for which they could supply a 'loose synonym' (State 4) from sets offering the more spaced learning opportunities; indeed, the sums of State 4 words prove broadly similar from children having read Sets 1, 2 and 5 alike –the maximum difference from reading any two sets amounts to 5 words (Sets 1 and 2), and the minimum to just 2 (Sets 5 and 2). Second, and with words *known* to the standard of VSAT States 6 and 5 comprising the bulk of gains from having read Set 5 in cases (**a**, $_{6+5+4,\ 5\&1}$) and (**b**, $_{6+5+4,\ 5\&2}$), the lexical competencies associated with these states come across as somewhat more responsive to a spacing effect than the particular competence State 4 seeks to capture. The sign test findings regarding State 4 learned words seem most interesting, arguably, in that the high p-values emerging from comparing *known* sums from a two-set combination reveal a lexical competence other than that State 6 captures that appears to exhibit insensitivity to presentation time manipulation (though see Section 5.8.1). This being so, the sensitivity of words in State 5 retains its status as the primary driving source of significantly ($p<0.05$) more word gains from the participant group reading one set of texts rather than another, when such cases arise. From a teacher's perspective, the arguably minimal knowledge sufficient for State 4 occupancy (see Chapter 3) suggests that State 4 insensitivity to spaced learning may have only minor pedagogical implications. The most obvious value of State 4 knowledge lies in little more than the assistance it affords in comprehending relatively simple written materials together with a role in enabling students to discount options among a set of distracters in a multiple-choice test. Beyond this, seems nothing other than a modest contribution to a child's general listening skills; one might suppose, for example, that a student able to define a word loosely would understand it receptively should s/he encounter it in a lesson or the playground. From a more theoretical perspective, State 4 insensitivity does, however, offer useful clarification regarding the particular lexical skills upon which manipulating target word presentation time has a likely impact. Given a single lexical knowledge continuum extending from zero familiarity with a word at one extreme, to full productive use at the other –see e.g. Waring (2000) for a discussion– these skills fall somewhere short of the capacity to supply words in well-formed clauses (State 6) and yet beyond whatever suffices to supply loose synonyms such as would justify word occupancy of VSAT State 4.

Conclusion 11: Defining *known* words as those that occupy States 6+5+4, the benefits of spaced learning reveal themselves disproportionately in gains of a very particular lexical competence –the capacity to provide good synonyms of formerly novel words.

Conclusion 12: More or less spaced learning impacts upon that range of lexical competence lying somewhere between, at the higher end, a productive capacity amounting to the ability to provide good synonyms of words, and the capacity to provide loose synonyms, at the lower.

5.9.3 Differences in target word presentation time on learning outcomes

The impact of presentation time upon *known* word sums with State 4 words included in the *known* totals, depends upon (a.) the receptivity of words in that state to a spacing effect, if receptive at all, and (b.) how substantial the sums of those words as a proportion of the *known* word total (on the latter point, as noted, the contribution seems 'small' or non-existent). That these totals appear unreceptive (Section 5.9.1) raises two points, when comparing word gains between sets of a two-set combinatoin. First, the difference in the number of signs (+ and - cases) that serves as input to both the 'general' and 'pairwise' testing procedures becomes rather less –that is, including State 4 words as *known* lowers the sum of paired cases from which children gained more target vocabulary from Set 5 (see footnote 92 for an explanation). Second, we no longer have a significant difference in *known* word sums under the 6+5+4 test of *knowing* from participants having read Sets 5 and 3, despite the respective sums proving significantly different with *known* words restricted to those occupying exclusively VSAT States 6+5 (case $\mathbf{c}_{,\,6+5,5\&3}$; see Section 5.8.3). In this particular instance (i.e. Sets 5 and 3), albeit not in cases ($\mathbf{a}_{,\,6+5+4,\,5\&1}$) and ($\mathbf{b}_{,\,6+5+4,\,5\&2}$), State 4 insensitivity appears to have lowered the total number of pairs of scores from which children gained more words from their Set 5 reading experience –indeed, doing so sufficiently to raise p-values above 0.05 and therefore leaving the null hypothesis firmly intact. Under the 6+5 test of *knowing*, 17 participants gained more words from Set 5 (compared to Set 3), while 12 did so under the 6+5+4 test.

As under the 6+5 test of *knowing*, the minimum addition to presentation time necessary for children to gain statistically more words from reading a set of a possible pairing remains 48 hours. Whether more impressive gains might have arisen from time additions lying between 24 and 48 hours (e.g. 30 hours, 35 hours etc.) remains unknown given the study's exclusive focus upon the effects of 24-hour 'stepped' presentation time increments. Including State 4 words in the *known* class sheds no additional light on whether intermediate differences in presentation time might ensure significantly more word learning (e.g. 12 hours, 8 hours etc.), or even the greater likelihood of learning beyond that from analyses restricting *known* words to occupants of States 6+5. Why, in case (**b**, $_{6+5+4,\ 5\&2}$), an addition to target word presentation time of 72 hours sufficed for children to gain significantly more words from Set 5 than Set 2, and yet proved insufficient for the participant group to gain significantly more from reading Sets 4 and 1 (p=0.134) –again a 72 hour difference– remains unknown.

5.9.4 Known words (6+5+4), and measures of gain; Research Question 1

Figure 5.22 presents a summary of the differences in *known* word totals from reading each set of a possible pair, the average number of words gained, and the percentage increase in target words participants learned from their reading experiences. From Set 5, for example, we see that the participant group gained 23 more words than from Set 1 (**a**, $_{6+5+4,\ 5\&1}$), amounting to an average of 0.83 additional words per child (from a possible 4) and a 33% increase over the Set 1 total. For case (**b**, $_{6+5+4,\ 5\&2}$) children learned an extra 20 words, corresponding to an increase of 27.7% and an average of 0.71 extra words from the more spaced (Set 5) learning opportunity. The sum of two-set combintions from which children gained additional words remained 'small,' however, amounting to just 2 (represented here by the data containing cells), or 20% of all possibilities (i.e. possible set pairings), leaving 80% of pairs (i.e. 8 out of 10) from which they failed to gain more words from the set offering the more spaced learning opportunity. A binomial sign test applied to these totals (8 and 2), zeros shared, identified no statistically significant word gain differences. Placed in numerical, as opposed to percentage, terms the extra *known* word sums might seem quite modest. From case (**b**, $_{6+5+4,\ 5\&2}$) the participant group gained on average less than one additional word from the more distributed word presentation of Set 5,[93] and less still from the same set in case (**a**, $_{6+5+4,\ 5\&1}$). The differences in *known* sums as a proportion of the total target words within a set (i.e. 112 words: 28 nouns + 28 verbs + 28 adjectives + 28 adverbs) appear a little more imposing, amounting to 17.8% in case (**b**, $_{6+5+4,\ 5\&2}$) and 20.5% in case (**a**, $_{6+5+4,\ 5\&1}$). Despite such arguably unexceptional gains,[94] the possibility nevertheless remains of substantial vocabulary development over the long term –i.e. several months or years– should children read texts designed to the specifications of either one or another set the investigation employed. Tentative projections of gains over extended time periods form the subject matter of Part 3. The minimum difference in target word presentation time that sufficed to induce significant differences in word gains proves quite large, at 72 hours i.e. three days (Figure 5.22). This addition did not always guarantee gains reliably attributed to spaced learning, however. As noted, participants failed, for example, to gain significantly more words from reading Set 4 than from Set 1 with *known* words defined as those that occupy States 6+5+4 or State 6+5.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					Difference in the sum of words gained: 23 Difference in the average number of words gained: 0.83 Percentage increase in words gained: 33%
Set 2					Difference in the sum of words gained: 20 Difference in the average number of words gained:0.72 Percentage increase in words gained: 27.7

Figure 5.22: Differences in the averages, *known* sums and percentages of target words gained (6+5+4).

[93] The theoretical maximum difference in the average of gains amounts to 4 which would arise were children to have gained no target words from reading one set of texts and all four target words from reading the other.
[94] The point comes from Hollander and Wolfe (1999).

5.10 General conclusions on differences in sums of learned words from reading the experimental sets

Sections (5.8-5.9) commented upon the proportions of *known* words among individual VSAT (i.e. State 6, State 5, and State 4) to provide more detailed insights into distributed learning outcomes than emerge from bare sums of words children *knew* to the standards of *known* that the study employs (i.e. occupancy of either State 6, States 6+5, or States 6+5+4). The current section draws together findings from these discrete, state-by-state, analyses to uncover general conclusions applicable to the three senses of word *knowing* the study acknowledges. The section seeks to uncover *common* circumstances under learning gains from a spacing effect will arise. The total words gained from reading any pairing of texts appears in Figure 5.23 for each of the three definitions of *known* under discussion.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1		6 (**2**) 6+5 (**2**) 6+5+4(3)	6 (**3**) 6+5 (4) 6+5+4 (12)	6 (0) 6+5 (14) 6+5+4 (11)	6 (4) 6+5 (20) 6+5+4 (23)
Set 2	6 (**2**) 6+5 (**2**) 6+5+4 (3)		6 (**1**) 6+5 (6) 6+5+4 (9)	6 (2) 6+5(16) 6+5+4 (8)	6 (6) 6+5 (22) 6+5+4 (20)
Set 3	6 (**3**) 6+5 (4) 6+5+4 (12)	6 (**1**) 6+5 (6) 6+5+4 (9)		6 (3) 6+5 (10) 6+5+4 (**1**)	6 (7) 6+5 (16) 6+5+4 (11)
Set 4	6 (0) 6+5 (14) 6+5+4 (11)	6 (2) 6+5 (16) 6+5+4 (8)	6 (3) 6+5 (10) 6+5+4 (**1**)		6 (4) 6+5 (6) 6+5+4 (12)
Set 5	6 (4) 6+5 (20) 6+5+4 (23)	6 (6) 6+5 (22) 6+5+4 (20)	6 (7) 6+5 (16) 6+5+4 (11)	6 (4) 6+5 (6) 6+5+4 (12)	

Figure 5.23: Differences in the sums of *known* words from reading alternative sets (figures in bold indicate a massed learning advantage).

Figure 5.23 clearly reveals the one 'directional' impact of presentation time manipulation. In all cases in which sign tests identified a significant difference in *known* word totals, whether of words differentiated by class or by VSAT state, children collectively typically derived the larger sum from the set offering the more spaced learning opportunity. Even so, the pairs of sets from which participants gained additional words from that offering the more massed learning amount to comparatively few – just 5 out of 30 (the figures in bold). This contrasts with the 24 pairs (i.e. 80%) from which higher gains arose under relatively spaced target word presentation. Where children did learn more words under the relatively massed learning condition, the gains proved arguably 'small,' amounting to between just 1 and 3 words. In contrast, from sets providing the more spaced learning condition, gains ranged from a low of 3 words to a high of 23. Collectively, the data in Figures 5.22 and 5.23 support the following conclusions:

<u>Conclusion 13 (General)</u>: In cases where children gained statistically significantly more words from reading one set of a pair, irrespective of the definition of *known*, the set providing the more distributed learning condition gave rise to the higher *known* word sum.

More generally:

<u>Conclusion 14 (General)</u>: For any set of texts of a pair (whether differences in known sums differed significantly or otherwise), the set affording the relatively massed presentation of target words rarely gave rise to more word gains than the set providing the more distributed presentation. Should children have gained more words from the relatively massed learning experience the difference in *known* word sums never proved statistically significant.

The likelihood of significant differences in sums of words occupying the same VSAT State i.e. 6 or 5 or 4 (see Figure 5.24) across sets from spaced learning opportunities arises from the vote-count/sign test findings, an arguably more sensitive measure of the 'big picture' than 'pairwise' tests limited to processing data from just two sets of interest. For VSAT State 6 words, children gained more vocabulary from the relatively spaced learning experience in 6 of 10 cases, and less in 3, with one tie. After assigning the tie to the N- cases (see Section 4.21) the test returns a p-value of 0.75, a figure consistent with the null hypothesis i.e. the no spaced learning effect conjecture (note, however, the aforementioned visual hint of an effect from the additions to State 6 sums with each 24-hour increment to target word presentation time exceeding 3 days –see Figure 5.23). For sums of State 4 words, an association between target word presentation time and gains appears more doubtful still, with the participant group learning the larger word sum from the relatively massed encounters in 5 out of the 10 possible two-set combinations, the vote-count/sign test returning a value of p=1.00. Together, the findings leave the apparent responsiveness of words familiar to the State 5 standard as the most substantial contributor to statistically significant differences in *known* words in all 6 cases in which these arise. This sensitivity to presentation time manipulation appears impressively robust, the vote-count/sign test returning a significant p=0.021 (Section 5.6.2), and 'pairwise' standard sign tests indicating four pairs of sets from which children gained significantly more words under the States 6+5 sense of *knowing* from their relatively spaced learning opportunities (i.e. **a**, $_{6+5,\ 5\&1}$; **b**, $_{6+5,5\&2}$; **c**,$_{6+5,5\&3}$; **d**, $_{6+5,4\&2}$). In the latter two cases binomial sign test findings imply the additional words familiar to the State 5 competence likely accounted for a disproportionately large sum (over 70%) of the total extra that children gained from their reading experiences (See Section 5.8.3).

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1		6 -2 5 0 4 5	6 -3 5 7 4 8	6 0 5 14 4 -3	6 4 5 16 4 3
Set 2	6 -2 5 0 4 5		6 -1 5 7 4 3	6 2 5 14 4 -8	6 6 5 16 4 -2
Set 3	6 -3 5 7 4 8	6 -1 5 7 4 3		6 3 5 7 4 -11	6 7 5 9 4 -5
Set 4	6 0 5 14 4 -2	6 2 5 14 4 -8	6 3 5 7 4 -11		6 4 5 2 4 6
Set 5	6 4 5 16 4 3	6 3 5 7 4 -11	6 7 5 9 4 -5	6 4 5 2 4 6	

Figure 5.24: Differences in *known* word sums by VSAT State (a figure preceded by a – sign indicates additional words from the set providing the more massed target word presentation).

Other observations from Figure 5.24 enjoy rather more 'tentative' support. That sums of words in both States 6 (controversially) and 5 respond to differences in target word presentation time does not imply comparable responsivity. The diminishing returns in gains of State 5 words stands out in Figure 5.24 with children gaining 7 extra words in this state from Set 4 over those from having read Set 3, and yet just 2 additions from reading Set 5 over those from reading Set 4. No less apparent, yet standing in marked contrast, are the more impressive State 6 gains with each 24-hour extension to target word presentation time beyond 72 hours. It seems possible, although as yet unproven, that presentation time exceeding 5 days would result in larger gains of words familiar to the State 6 standard (from set to set comparisons), and perhaps lesser gains of those familiar to the standard of State 5. Also plausible is an earlier onset of a decreasing spaced learning advantage with '24-hour' extensions of presentation time beyond five days (72 hours), if limiting *known* words to those occupying State 5 as opposed to States 6 and 5 combined. The permutations of *actual* presentation time and *difference* in presentation time that suffice for significantly more word learning to the State 6 or State 5 competence remain largely unclear (the issue extends beyond the narrow ambit of the Research Questions), however, as does the number of 24-hour additions to presentation time beyond five days that might result in children ceasing to gain additional words to the VSAT State 6 standard. To investigate such issues would call for sets of texts exposing children to target words over more than one school week (120-hour) period. It does seem clear, nevertheless, given that States 6 and/or 5 constitute elements of all *known* word definitions (i.e. 6, 6+5 or 6+5+4), that '24 hour' incremental increases to target word presentation time will differentially impact both the number of *known* words, and the depth to which children *know* them, depending upon the definition of *knowing* we acknowledge. As we have seen, a difference in target word presentation time (and actual presentation time) resulting in significantly higher ($p<0.05$) *known* totals under the 6+5 test of *knowing* may not ensure the largest gains under the 6+5+4 notion given the same pair of sets. It appears possible, also, and perhaps even likely, that the effects of spacing will manifest themselves in more or less substantial gains depending upon which of any reasonable notions of word *knowing* one subscribes to.

That the study finds children learned most target words from reading Set 5 (i.e. from novel word exposure extending over five days) under all definitions of *known* does not mean such gains necessarily prove more durable than those from reading another. The data in Figure 5.24 supplies no indication of memory decay, a subject that has attracted little research generally, and less still in regard to vocabulary gains arising exclusively from RR sessions. It seems implausible that words children learned from a set of texts became permanent additions to a child's lexicon given the absence of reinforcement once the study ended, given the absence of encounters with target words outside of the classroom. The question remains, also, whether children would retain or lose words *known* to the criterion of one VSAT State (e.g. State 6, as opposed to State 5) more readily than those in another. Quite possibly, rate of forgetting might have proven substantial, as it did among Waring and Takaki's (2003) participants in a study exploring word gain duration from reading experiences. The results from the present investigation reveal only sums of words children retained as measured by tests typically undertaken within a few minutes of having completed the final text of a set.

Above all, the relationship between target word presentation time and likelihood of learning emerges as both complex and 'messy' with gains arising from as yet little understood factors and the subtle, conjunctive, interactions between them. Certainties seem few and far between. Why, for example, with *known* words limited to those in States 6+5, does a 48-hour addition to presentation suffice for participants to learn significantly ($p<0.05$) more target words from Set 4 than Set 2 in case (**d**, $_{6+5,\ 4\&2}$), but not for children to gain more words ($p<0.05$) from Set 3 than Set 1. And why, if we define *known* words as occupants of VSAT States 6+5+4, does a 72-hour difference in target word presentation time suffice for participants to gain significantly more vocabulary from Set 5 than from Set 2 (case **b**, $_{6+5+4,\ 5\&2}$), yet fail to do so from having read Set 4 rather than Set 1? Even for the same pair of sets, an addition to presentation time associated with a statistically significantly ($p<0.05$) word gain, under one definition of *known* does not necessarily imply such gains under another. Children gained significantly more words from reading Set 5 over Set 3 under the 6+5 notion of *knowing* and yet failed to do so if *known* denotes words occupying States 6+5+4.

For the same notion of *knowing*, the data from the current investigation failed to identify any presentation time addition (in hours) that would ensure (i.e. guarantee) children collectively gained additional words over those from a set offering a more massed learning experience. That the participant group learned significantly ($p<0.05$) more words from Set 5 than Set 1 in case (**a**, $_{6+5,\ 5\&1}$) provides no assurance that the same four day difference in target word presentation time would have sufficed had the group encountered target words over 10 days as opposed to 6, or 11 as opposed to 7. One fact stands out, however: Suggestions that presentation time extensions will deliver statistically significant differences in learning outcomes only becomes meaningful if we also qualify what the term *known* implies. As a general conclusion, the present research attributes significantly more words from a set of a pair to the combined effects of three factors: (1.) what is a *known* word, (2.) the *addition* to presentation time a set affords –i.e. the extra hours a Set presented its target words compared to the other, and (3.) the actual *time* (as opposed to additional time) over which readers encountered target words from reading one or other set of the two.

The minimum addition to presentation time below which significant differences in word learning fail to arise remains unclear. Forty-eight 48 hours sufficed for children to gain significantly more words from a set of a pair in case **c**$_{,6+5,5\&3}$, for example, and also case **d**, $_{6+5,\ 4\&2}$, but this still leaves undetermined whether time intervals of between 24 and 48 hours might have accounted for more, or equal, learning ($p<0.05$), and how *known* word sums might have varied depending upon the notion of word *knowing* we choose to acknowledge. The same uncertainty holds for words occupying the individual VSAT states (i.e. 6 or 5 or 4) that comprise the 'building blocks' of the alternative notions of *knowing* the study adopts. The minimum addition to presentation time below which a spacing effect fails to express itself in significantly more words occupying State 5, for example, may not necessarily apply to words in States 6 or 4, even assuming the unlikely event that sums in the latter states indeed exhibit sensitivity to spaced learning opportunities.

Conclusion 15 (General): For texts designed to the specifications of Sets 1 to 5, the minimum additional time period over which one set of a pair need present its embedded target words before a statistically significant ($p<0.05$)

difference in *known* word sums is observed was 48 hours.

Conclusion 16 (General): For the sum of words which occupy VSAT State 5 to differ significantly from reading one or other set of texts of a pair, the minimum additional presentation time amounted to 48 hours.

Conclusion 17 (General): For the same definition of *known* word, the greater the additional time over which a set of texts presented its target vocabulary, the more likely we are to observe a statistically significant difference in *known* sums (caveat: Under the State 6 based test of knowing, even 4 additional days proved insufficient).

Conclusion 18 (General): Whether a particular difference in the time period over which a set of a pair (of texts) presented its target words resulted in statistically significant differences in learning outcomes depended upon which of three definitions of *known* word one acknowledges.

The lexical competencies upon which spaced target word presentations impact now seem reasonably clear. The study finds strong evidence that word presentation time affects the sums of words children could define to a native-like standard (i.e. the low level productive competence State 5 captures), and some indication (albeit little more than a mild hint, perhaps) that intervals between encounters determine how many words children could supply in grammatically and syntactically well-formed clauses (i.e. the VSAT State 6 standard). The sums of words for which the participant group could supply a loose synonym (State 4) proved highly unresponsive to presentation time manipulation (though see Section 5.15), whether the comparison involved the vote-count/sign test or the 'pairwise' testing process. Viewing such findings in the light of 'single,' as opposed to 'multi,' continuum representations of lexical proficiency (see e.g. Waring, 2000 for a discussion), the study locates the effects of spaced learning to a zone of lexical competence with an upper boundary lying short of native-speaker productive capacity (recall that the grammatical test for occupancy of State 6 refers to communicative competence), and a lower amounting to the ability to supply a relatively poor target word synonym (State 4). That the participant group nevertheless collectively learned some words to the standard of States 6 and 4 from each set of texts would at least suggest, however, the likelihood of further gains had they continued to read texts designed to the specifications of any set, or sets, of texts, that the study employed. Of the VSAT states 6, 5 and 4, sums in State 5 would most likely differ significantly ($p<0.05$) in response to stepped (24 hour) increases in target word presentation time additions of up to 96 hours, the maximum extension the study examines. Whether children might have learnt significantly more from encountering the 12 occurrences of each word had a set (of a pair) offered in excess of 96 hours of distributed learning remains pure speculation, as does the likely learning outcome had participants encountered each target word on more, or less, than the 12 occasions that the texts provided. Answering such questions calls for quantifying vocabulary uptake over periods exceeding five days, and a methodology that addresses the practical difficulty of a 5-day school week and 2-day weekend break.

Conclusion 19 (General): The effects of manipulating the time interval between reencounters with the same novel word primarily impact upon the sums of words for which children could supply a 'good' definition.

Given so few generally applicable findings, the study supports a range of apparently contradictory positions, yet each defensible from the evidence of word gains from the reading sessions:

1. Spaced learning *does not* result in statistically significantly ($p<0.05$) more word gains; this captures the view of those for whom a *known* word amounts to an occupant of VSAT State 6 (that is, a word which a child could supply in a semantically and syntactically correct clause).

2. Spaced learning *does* give rise to statistically significantly ($p<0.05$) more word gains in *known* words (where known denotes occupancy of VSAT States 6+5); examples are cases (**a**, $_{6+5,\ 5\&1}$); (**b**, $_{6+5,\ 5\&2}$); (**c**, $_{6+5,\ 5\&2}$); and (**d**, $_{6+5,\ 4\&2}$).

3. Spaced learning *does not* account for statistically significantly ($p<0.05$) more *known* words (to the standard of VSAT States 6+5), a finding that holds true for any pair of sets other than cited in 2 above. For example, participants failed to gain more words overall from reading Set 4 than Set 1.

These observations caution against overarching claims (see Dempster, 1988; Krashen, 2004) that distributed presentation necessarily proves more 'learning conducive' than massed *per se,* suggesting, rather, a more nuanced, qualified, endorsement (Nation, 2000). The (possible) insensitivity of State 6 sums to spacing, in particular, would suggest that textual adaptation to control time intervals between novel word encounters will likely prove ineffectual for purposes of developing ‘high level’ productive word knowledge of the sort that underpins effective writing and speaking. On the other hand, adaptation becomes rather more productive, at least from a teacher’s perspective, assuming the less ambitious goal of expanding a child’s receptive language competence –i.e. raising the sum of words in State 5. Even so, and despite clear evidence for State 5 sums’ apparent responsiveness to intervals between target word encounters (see Section 5.8.2), generalizations are still apt to mislead. Children as a group did not necessarily gain statistically more ($p<0.05$) State 5 words from the same pair of sets under alternative definitions of *known* word. Nor did a presentation time difference (i.e. an additional 24, 48, 72 hours etc.) associated with significantly larger State 5 sums from reading a set of a pair necessarily give rise to a statistically significant difference for an alternative pair (Section 5.8.3). The study suggests a complex, subtle, yet unestablished, relationship between State *5* totals and spaced learning opportunities that calls for further research –research all the more relevant given the pedagogical importance of children understanding words to this particular *known* standard. This leaves the factors that predict significantly more words familiar to the VSAT State 5 competence as those predictive of differences in *known* word sums generally (i.e. the sum of words in States 6+5, or 6+5+4), namely: (1.) the *actual* times readers expend completing the sets of texts of a pair; this measured in days, and ranging from between one and five, and (2.) the difference in presentation time –measured in 24 hour increments– over which one set of texts presents its complement of target words compared to that of the another. Both factors interact in some fashion. For the *same* 72 hour presentation time difference, the participant group gained statistically more words from Set 4 than Set 2 (expended reading time: 4 days, and 2 days), with *known* words as those in States 6 of 5, and yet failed to do so from reading Sets 3 and Set 1 (expended reading time: 4 days and 1 day, respectively), despite the same 72 hour presentation time difference.

<u>Conclusion 20</u> (General): For any pair of sets such that a significant difference was observed in the sums of *known* words from reading one or the other, given the particular *definition* of *known,* those same sums miss significance ($p<0.05$) under one or more alternative *definitions*.

Breaking down the definitions of *known* word into constituent VSAT states identified two possible mechanisms from which statistical differences in *known* sums from reading a set of texts of a pair might arise. Drawing on cases (**a**,$_{6+5,5\&1}$), (**b**,$_{6+5,5\&2}$), and (**c**,$_{6+5,5\&3}$) the study suggests a significant difference could occur from:

(1.) Participants learning more words 'generally' from the set of a pair affording the more spaced learning opportunity –that is, the participant group failed to gain statistically more in any one VSAT state compared to those occupying that same state from the other set. Case **c**,$_{6+5, 5\&3}$ arguably illustrates this possibility.

(2.) A disproportionate increase in the sum of words that occupy a single state –invariably State 5 (see, for example, cases **a**, $_{6+5\ 5\&1}$, and **b**, $_{6+5, 5\&2}$) – that 'noticeably' exceeded gains in other states from which the definition of *known* word derives.

Given the apparent immunity of sums of words in State 4 (and possibly State 6) to more or less spaced presentations this prompts the following general observation:

<u>Conclusion 21 (General)</u>: A significant difference in the sums of words in State 5 could (but did not always) represent the primary source of the additional *known* words from reading a set of texts of a pair.

Figure 5.25 displays the totality of pairs of sets associated with significantly more learning for each definition of *known,* along with measures of those differences in terms of averages, sums and percentages. Blank cells at a row/column intersection denote a pair of sets from which children failed to gain statistically ($p=0.05$) more words from reading one set of the two. The Figure reveals just how commonly a spaced learning advantage truly arose, but also the numerical limits of learned word sums. Of the total 30 pairs of sets (10 pairs for each sense of *knowing*) in only 6 (i.e. 20 % of cases) did additional target word presentation time account for children collectively gaining statistically significantly ($p<0.05$) more words from a relatively spaced learning opportunity. In all but one case (the pair of Sets 4 and 2) the participant group gained the larger total from Set 5, hinting perhaps at more substantial gains still had target word encounters occurred over a longer presentation time than the 5 days that the set provided. The upper limit to gains stands as the 23 additional words children *knew* from Set 5 as opposed to Set 1, an increase of 33%. The least substantial gain, yet statistically significant nonetheless (pairwise binomial sign testing -zeros apportioned), were the 16 additional words children learned from Set 5 over those from reading Set 3 (a 29% increase) with *known* words restricted to occupants of States 6+5. Under the 6+5 test of *knowing,* children gained significantly ($p<0.05$) more words from just 4 pairs of sets ('pairwise' testing), as opposed to 2 under the 6+5+4 test. A significant difference in *known* word sums under one notion of *knowing,* however, did not necessarily signify a significant difference under another. While children (as a group) always gained significantly ($p<0.05$) more words under the States 6+5 test of *known* should they have likewise achieved this under the States 6+5+4 definition, the opposite does not always hold true. Figure 5.25 indicates two cases where a disparity in learned word sums (namely, from reading Sets 4 and 2, and Sets 5 and 3) no longer persists under a notion of *knowi*ng that includes words in States 6, 5 and 4. As noted, the loss of a significant difference observed under the stricter (6+5) test plausibly arises from the insensitivity of State 4 sums to differences in target word presentation time. A possible explanatory mechanism is described in footnote 92. The inflationary effect on p-values from including State 4 words in the *known* word total seem all the more powerful to the extent such words comprise a higher proportion of the sum that children *knew* to the States 6+5+4 standard.[95]

	Set 1	Set 2	Set 3	Set 4	Set 5	
Set 1					(6+5), Sum:20 Av:0.72 Inc.: 39.2 %	(6+5+4) Sum: 23 Av: 0.83 Inc:33%
Set 2				(6+5) Sum:16 Av:0.57 Inc.:32.6%	(6+5) Sum:22 Av:0.79 Inc.: 44.8%	6+5+4 Sum: 20 Av:0.72 Inc.:27.7%
Set 3					(6+5) Sum:16 Av: 0.58 Inc.: 29%	

Figure 5.25: *Known* (6+5+4) words (*Sum=additional words from text offering the more distributed word presentation; Av= difference in the average number of words learned from reading the sets of a pair; Inc.= percentage increase in known words attributed to a spacing effect).*

[95] To take the case of Sets 5 and 3, the effect of adding State 4 words to the known totals is to reduce the number of cases in a child gained more from Set 5 than 3 from 13 to 7.

<u>Conclusion 22 (General):</u> Of the six pairs of sets where children gained statistically significantly more words from reading one or the other, the set providing the higher known total was Set 5 (i.e., the set providing the most distributed learning opportunity) in five of these cases.

<u>Conclusion 23 (General)</u>: Broadening the category of *known* words to include those in State 4 reduces the number of cases in which a child gained more words from the set affording the more spaced learning opportunity.

<u>Conclusion 24</u> (General): For a pair of sets, the maximum difference in the average number of words a child gained from reading that set from which s/he learned most amounted to an extra 0.82 words (from the possible 4). This corresponding to a 23 percent gain in his/her known word total.

<u>Conclusion 25 (General)</u>: For any pair of sets, the minimum difference in the average number of words a child gained from the set of a pair associated with significantly more *known* words was 0.28.

5.11 **Research Question 2: Statistical significance and between-sets analysis by word class**

The Research Question asks:

> *How significant (statistically and pedagogically) are differences in the sums of novel words of the four content word classes (nouns, verbs, adjectives and adverbs) that child readers gain from encountering those words of a particular class under more, or less, distributed learning conditions?*

Research Question 2 builds upon Research Question 1, moving the discussion beyond aggregate *known* word totals to address two further issues relevant to understanding spaced learning within the context of school-based RR: (1.) Whether the sums of words children gained of each lexical class differed significantly depending upon the time over which word presentations occurred (i.e. did spaced learning result in more gains of a specific word type?); and (2.) How substantially any such 'class gain' contributes to the gross total of *known* words from reading that set of a pair from which most learning arose? (see Part 1).

Data presentation follows the same format employed for Research Question 1, with findings in regard to each notion of *known* word (a. State 6; b. States 6+5 and c. States 6+5+4) receiving attention in turn.

5.11.1 *Known* words as those that occupy VSAT State 6 (Research Question 2)

The subject warrants only brief mention given McNemar findings (Section 5.4.2) that failed to reveal cases of statistically significant ($p<0.05$) differences in *known* sums of any class irrespective of possible set pairing. Indeed, testing identified several instances in which the participant group learned rather more words from the relatively massed target word presentation a set of a pair afforded (see Section 5.8.3 and Table 5.1). Children collectively gained more nouns from Set 1, for example, than from Sets 2 or 3, more verbs from Set 2 than Set 3, and an equal sum of verbs from Set 2 as from Set 4. The data for adjectives reveals more learning from Set 1 (highly massed) than Set 5, while sums of *known* adverbs proved identical for Sets 1 to 4 (children gained a total of just 3 words from each) and only marginally higher from reading Set 5 over those from having read Set 1. The vote-count/sign test for differences in learned word sums of any one class (N, V, Adj, or Adv) across the five sets of texts affirms the null hypothesis position, the test supplying the following values for the classes examined: nouns ($p=1.00$), verbs ($p=0.34$), adjectives ($p=1.00$), and adverbs ($p=0.34$).

<u>Conclusion 26</u>: With *known* words defined as those occupying State 6, participants did not gain statistically significantly more words of any one lexical class (noun, verb, adjective or adverb) from reading one or other of any two sets of a possible pair.

5.12 *Known* words as those occupying VSAT States 6+5 (Research Question 2)

With the class of known words broadened to include occupants of States 6+5, significant differences in *known* word totals (words undifferentiated by class) arise in 4 cases:

1. (**a**, $_{6+5,\ 5\&1}$): i.e., from reading Sets 5 and 1 ($p=0.01$)
2. (**b**, $_{6+5,\ 5\&2}$): i.e., from reading Sets 5 and 2 ($p=0.01$)
3. (**c**, $_{6+5,\ 5\&3}$): i.e., from reading Sets 5 and 3 ($p=0.02$)
4. (**d**, $_{6+5,\ 4\&2}$): i.e., from reading Sets 4 and 2 ($p=0.04$)

5.12.1. Observations on cases

Of these cases, only from (**b**, $_{6+5,\ 5\&2}$) and (**c**, $_{6+5,\ 5\&3}$) did children gain significantly more words of any single class –nouns in each instance with the larger sum associated with reading Set 5. In case (**b**, 6+5, 5&2), the gain amounts to 10 extra nouns over those from reading Set 2 (McNemar; p=0.002), corresponding to an addition of 58%. The contribution these extra nouns make to the total additional words gained (i.e. the sum of nouns, adverbs, adjectives and adverbs) appears substantial, comprising 45% of that sum. Verbs proved the second largest contributor comprising 27%, followed by adjectives (18%), and then adverbs (9%). Regarding case (**c**, 6+5, 5&3), participants gained 8 extra nouns from the more spaced learning Set 5 afforded, an increase of 42% over the total from Set 3 (McNemar, p=0.008), the gain accounting for 50% of the additional words overall from having read Set 5 texts. The two extra verbs contributed 12% of the gross gain, the 6 adjectives 37%, and the adverbs zero.

<u>Conclusion 27</u>: From reading Sets 5 and 2, and Sets 5 and 3, children gained significantly (p<0.05) more nouns from the set of the pair offering the more spaced presentation time.

<u>Conclusion 28</u>: The significant difference in nouns in cases (b, 6+5, 5&2), and (c, 6+5, 5&3) make a relatively substantial percentage contribution (45% and 50% respectively) to the sum of additional words children gained from more distributed target word presentation.

A McNemar 'pairwise' test on *known* sums from the participant group having read Sets 4 and 2 (case **d**, 6+5, 4&2) indicated the respective noun totals (the respective *known* noun sums form Sets 4 and 2) missed significance comfortably (p=0.18), as did the *known* word sums of the remaining three classes. Of the additional words (i.e. total undifferentiated by lexical class) gained from reading Set 4 over Set 2, nouns contributed 31% of that sum, the same percentage contribution as from verbs. Adjectives made up 25% of the gross gain, and adverbs just 12%. Adverbs aside, then, no single word class contributed notably more (or less) substantially to the additional *known* total than did any other. Both case (**d**, 6+5, 4 &2), and case (**a**, 6+5, 5&1) illustrate one of two manners by which extra learned words arising from spaced learning could account for significant differences in gross (undifferentiated by class) *known* word totals –namely, a general increase of all four word types stemming from the more spaced learning a set provided, as opposed to disproportionate additions to that total from words of any one class or classes. [96] An illustration of the second manner of contribution –a disproportionate addition from *known* words of a single lexical category, or categories– comes from Cases (**c**, $_{6+5,\ 5\&3}$) and (**b**, $_{6+5,\ 5\&2}$). Here we indeed see a relatively substantial contribution to the additional *known* word sum (i.e. *known* words undifferentiated by class) from specific lexical categories; in these cases, nouns, as mentioned (p. 196). All this seems to mirror the circumstances noted earlier (Section 5.9.1) under which statistically more words occupying a single VSAT state contribute to statistically significant differences in gross (i.e. undifferentiated by lexical class) *known* word sums from reading a set of a pair: either, a disproportionate contribution to that sum from, specifically, the additional State 5 words or, alternatively, a more general increase in words familiar to the standards of States 6, 5 and 4.

Case (**a**, $_{6+5,\ 5\&1}$) highlights the difficulty of inferring statistical significance without regard to the totality of evidence bearing upon null hypothesis validity. McNemar 'pairwise' tests on noun sums returned a value of $p=0.063$, a figure only marginally exceeding the 5% cut off (alpha, $p<0.05$) marking the conventional boundary for null hypothesis rejection. Strictly speaking, the null hypothesis remains firmly intact. On the other hand, we have compelling evidence for a spacing effect contribution to differences in learned noun totals from the pair of Sets 5 and 2 (**b**, $_{6+5,\ 5\&2,}$ above) and Sets 5 and 3 (**c**, $_{6+5,\ 5\&3}$). Can one reasonably suppose, in the light of this, that a spacing effect so demonstrable in these latter cases no longer applies in Case (**a**, $_{6+5,\ 5\&1}$)? Cases (**b**, $_{6+5,\ 5\&2}$) and (**a**, $_{6+5,\ 5\&1}$) seem informative for resolving the dilemma since the difference in target word presentation time from reading either pair (i.e. Sets 5 and 2, or Sets 5 and 1) amounts to just 24 hours –that is, Set 1 exposes readers to target words over a single day, and Set 2, two days. Ruling out a spacing effect contributing to the additional nouns in case (**a**, $_{6+5,\ 5\&1}$) might, now seem unduly severe, in effect reducing determination of statistical significance to a simple 'yes/no' response when the issue more properly concerns 'degree' –namely, whether the very real effect observable in cases (**b**, $_{6+5,\ 5\&2}$) and (**c**, $_{6+5,\ 5\&3}$) remains operative in case (**a**, $_{6+5,\ 5\&1}$), albeit rather less obviously so. To acknowledge an effect –i.e. to accept that the distributed

[96] In case (d, 6+5, 4&2), children did not, for example, gain statistically significantly ($p<0.05$) more nouns, or verbs, or adjectives, or adverbs from reading Set 4 as opposed to Set 2 but did gain statistically more words overall from the Set 4 reading experience.

noun encounters from reading Set 5 explains the extra nouns– means rejecting the ‘no spacing effect’ position on the grounds that relevant evidence should comprise rather more than a single McNemar finding. Given that the additional nouns in (**b**, $_{6+5,\ 5\&2}$) indeed arise from a spaced learning advantage (the p-value of 0.002 proved highly significant, after all), then abandoning the null hypothesis in case (**a**, $_{6+5,\ 5\&1}$) becomes rather more reasonable –arguably the rational outcome of a discriminating approach to data interpretation that embraces a broad view of the available data. Whether one can prudently, or should, propose an operative spacing effect in case (**a**, $_{6+5,\ 5\&1}$) raises a somewhat different issue, the answer to which depends upon the importance attached to avoiding a Type 1 error, weighed against the likelihood such an error arises at all –whether, that is, a 1 in 16 chance of a finding which would arise were the null hypothesis indeed correct (as implied by p=0.063 for case **a**, $_{6+5,\ 5\&1}$) permits the inference that spaced learning indeed made no contribution to gains. That the extra nouns from reading Set 5 over those from Set 1 amount to less than the additional verbs (sums of *known* verbs from reading all sets of texts miss significance comfortably; p=0.118), and only marginally exceed the gain in adjectives only adds to the interpretive difficulty. Such findings underline the admittedly tentative grounds sustaining the case for spaced learning as the source of *known* noun (sum) differences in the (**a**, $_{6+5,\ 5\&1}$) case. They also affirm that maintaining the null hypothesis remains an obvious, defensible, and arguably the more prudent option.

If, on the sum of evidence, we accept the null hypothesis and so deny that spaced learning accounts for the additional nouns from reading Set 5 as opposed to Set 1in case (**a**, $_{6+5,\ 5\&1}$) then here stands another example, along with case (**d**, $_{6+5,\ 4\&2}$), in which a significant difference in gross, undifferentiated, *known* word totals arise primarily from a general increase in gains of all four word types. This general gain reveals itself in the data on learning outcomes. Of the 20 extra words children gained from reading Set 5 in case (**a**, $_{6+5,\ 5\&1}$), the sums of each class indeed proved more or less comparable: 5 were nouns, 7 were verbs, 4 were adjectives and 4 were adverbs. Case (**a**, $_{6+5,\ 5\&1}$) serves to illustrate that significant, or near significant, p-values need not imply comparable pedagogical importance. Children could gain few extra words from a set of texts despite ‘pairwise’ or ‘general’ testing that returned p-values far below the critical alpha of 0.05.

<u>Conclusion 29</u>: Case a, 6+5, 5&1 stands as an example of a general increase in words gained from reading Set 5 as opposed to disproportionately more of any one lexical class.

5.12.2 Observations on presentation time

Part 1 (Sections 5.5-5.5.3) established that a difference in target word presentation time from which readers gained significantly ($p<0.05$) more *known* words from a set of a pair might not result in such additional learning given an alternative pair. We now see this applies likewise at the level of word class. The 48-hour addition to target word presentation associated with statistically more noun learning from having read Set 5 compared to that from Set 3 –*known* words defined as those occupying VSAT States 6 and 5– did not result in statistically ($p<0.05$) more nouns learned from reading Set 4 than from Set 2 (case **d**, 6+5, 4&2; p=0.180). Nor, indeed, did the participant group gain more nouns from Set 4 rather than Set 1 (p=1.000) despite that the same 3 day addition to presentation time reasonably explained a significant difference in learned nouns in case (**b**, 6+5, 5&2; p=0.002), albeit prior to apportioning zeros). The factors that predict statistical differences in learning outcomes under the State 6+5 notion of *knowing*, therefore, would seem to correspond to those predictive of gross *known* sum totals generally (i.e. sums undifferentiated by class): namely, (1.) the difference in presentation time (in hours) between the sets of a pair, and (2) the actual time over which children encountered the target vocabulary from those same sets. Neither factor suffices by itself to explain gains, however. Both Sets 5 and 2 (i.e. case b, 6+5, 5&2) and Sets 1 and 4, for example, presented their target words over the same time period –a span of seventy-two hours (i.e. zero presentation time difference) – and yet only in the former case (Sets 5 and 2) did participants gain significantly more words *overall* from the set offering the relatively spaced learning opportunity (for the Set 5 and 2 comparison, p= 0.01, as opposed to the non-significant p=1.00 for Sets 4 and 1 comparison). The *actual* presentation times, in contrast, differed markedly; on the one hand 120 hours and 48 hours (case **b**, 6+5, 5&2), and on the other, 24 hours (Set 1) and 96 hours (Set 4).

<u>Conclusion 30</u>: A statistically significant difference in the sums of known words of a particular class from reading a set of a pair does not necessarily imply a statistically significant difference in words of the same class from reading

an alternative pair of sets, even though the additional time over which target words appeared remained the same in each case.

The likelihood of observing a spaced learning advantage on *known* noun sums under the State 6+5 test of *knowing* corresponds to the ratio of the number of pairs of sets from which children gained statistically more words of that class (nouns), and the number from which they did not. The respective totals help place the pervasiveness of spaced learning into some kind of overall perspective. Of the ten possible pairs of sets, only from two did children gain significantly (p=0.05) more nouns from the relatively spaced learning condition –the aforementioned cases (**b**, $_{6+5,\ 5\&2}$) and (**c**, $_{6+5,\ 5\&3}$) described above. This leaves 8 pairs of sets (i.e. 80%) from which they failed to do so. The chance that either a child, or the participant group, would gain more words of any class *other than* nouns from reading a set of a pair emerges lower still. McNemar findings in Section 5.5.3 reported no significant differences in learning outcomes for the same class for any of the 10 pairs of sets examined.

<u>Conclusion 31</u>: With *known* words defined as those in States 6+5, a statistically significant difference in words of a particular class was observed in just 2 pairs of sets from a total 40 in which such a difference might potentially have arisen.

<u>Conclusion 32</u>: Of the 4 pairs of sets of texts from which children gained more words *overall* (where *overall* refers to the sum of nouns, verbs, adjectives and adverbs) from reading a set of a pair, in two of those cases the gain is reasonably attributable to a disproportionate sum of additional *known* nouns as opposed to words of other classes.

5.12.3 Distribution of *known* words among states

The distribution of *known* nouns among VSAT States 6 and 5 for cases (**b**, $_{6+5,\ 5\&2}$) and (**c**, $_{6+5,\ 5\&3}$) appears in Figure 5.26 (below). The State 6 sums received attention previously in Section 5.11.1 that cited evidence affirming the ‘general’ and ‘pairwise’ null hypotheses that deny a spaced learning advantage (Section 5.11.1). Findings from applying the McNemar ‘pairwise’ test to State 5 *known* noun totals, however, yield two very different results. For case (**b**, $_{6+5,\ 5\&2}$), the sums of State 5 nouns miss significance only narrowly (p=0.065; zeros not apportioned), raising the same interpretive difficulties encountered when discussing noun totals undifferentiated by state from children reading

Sets 5 and 1 (see Section 5.12.1). The null hypothesis, however, appears rather less tenable in case (**b**, $_{6+5,\ 5\&2}$) than in the aforementioned Sets 1 and 5 analysis given the higher probability revealed from McNemar tests (1 in 15.4, as opposed to 1 in 16 for case **a,** $_{6+5,\ 5\&1}$) that the difference in noun sums arise from chance. The suggestion –or 'hint'– that the spacing may nevertheless explain the disparity in State 5 *known* nouns in case (**b**, $_{6+5,\ 5\&2}$) comes from the vote-count/sign test which returned a p-value of 0.02, taking as input the 9 pairs of sets (the N+ observations, from which children gained more nouns to the State 5 standard), and the one (N - observation) from which they did not. The probability of this disparity (9 and 1), assuming a valid null hypothesis conjecture, only fractionally exceeds 1 in 50, a ratio falling well below the 1 in 20 traditionally sufficing for null hypothesis rejection. The implication is, arguably, a spacing effect that McNemar 'pairwise' tests failed to detect given the small samples of nouns available for comparison; the implied conclusion of a Type 2 error, however, remains tentative given the limited data available for analysis (see footnote 89).

In case (**c**, $_{6+5,\ 5\&3}$) the difference of three extra State 5 nouns from Set 5 comfortably misses statistical significance ('pairwise' McNemar test; p=0.607). This suggests, given the 'limited' evidence of a spaced learning advantage in regard to State 6 *known* word sums (Section 5.11.1), the significant difference in *known* nouns overall (i.e. the total words of this class occupying States 6 and 5) reasonably arises from more words familiar to the standard of both States 6 and 5 as opposed to State 5 alone. In the marginal case of (**a**, $_{6+5,\ 5\&1}$), for which the evidence of spacing as an explanation of differences in *known* State 5 noun sums remains contentious (p=0.063), the respective noun totals in State 5 from having read Sets 5 and 1 miss significance by a large margin (McNemar, p=0.388). Assuming children (collectively) indeed *knew* statistically significantly more nouns under the 6+5 test of *knowing* in case (**a**, $_{6+5,\ 5\&1}$) this, again, most plausibly arises from a 'general' spacing effect operative upon words familiar to the States 5 and 6 standards as opposed to an effect disproportionately acting upon sums in one State or the other.

<u>Conclusion 33</u>: With *known* words defined as those in States 6+5, children gained significantly more nouns from reading Set 5 than they did from reading either Set 2 (case b, $_{6+5,\ 5\&2}$**) or Set 3 (case c,** $_{6+5,\ 5\&3}$**).**

<u>Conclusion 34</u>: Regarding Sets 5 and 3 (case c, $_{6+5,\ 5\&3}$), the source of the statistical difference in *known* nouns appears a likely general increase in nouns familiar to both the State 6 and State 5 criterion of *knowing* arising from the more distributed learning Set 5 afforded.

<u>Conclusion 35</u>: The source of the extra nouns from reading Set 5 over those from reading Set 2 (b,$_{6+5,\ 5\&2}$) stems, arguably, from a disproportionately large sum of nouns familiar to the standard of VSAT State 5 arising from the more distributed learning associated with having read Set 5.

More generally:

<u>Conclusion 36</u>: The time intervals between participants' encounters with the same target noun affect not just whether they *knew* significantly more from reading a set of a pair, but also the relative proportion of *known* nouns among the two States (5 and 6) that define the *known* word class.

<u>Conclusion 37</u>: In all four cases where the difference in total *known* words (States 6+5) was observed, there emerges an excess (or equal number) of words of all lexical classes in the set associated with most word learning.

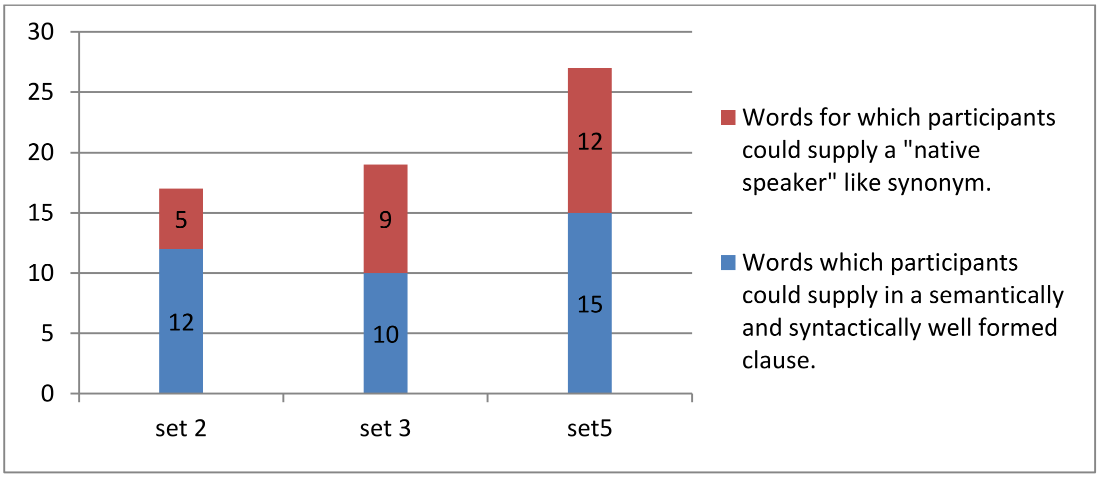


Figure 5.26: Totals of nouns *known* to the standard of VSAT States 6 and 5.

McNemar test findings (Part 1) indicated that statistically significant differences in *known* noun sums (the aggregate of those in States 6 or 5) from a set of a pair do not arise other than from instances in which children gained a significant difference in gross *known* word sums (i.e. the total words *known* undifferentiated by class) from that same pair. Given this observation, the circumstances under which the participant group gained more words of a class now become reasonably clear: Apart from readers having to have gained significantly ($p<0.05$) more words overall (i.e. a larger gross sum of target words irrespective of class) from a set relative to gains from another, two further conditions need apply. First, the sets must necessarily include Set 5 –that set offering the most distributed target word presentation. Second, the remaining set will consist of either Set 3 or Set 2. The small numbers of nouns children gained from reading the experimental texts precludes more definitive conclusions.

Conclusion 38: Only if target word presentation occurred over 120 hours (i.e. from reading Set 5) did children gain more nouns over those from reading another set.

The minimum difference in target word presentation time that sufficed for children to gain significantly ($p<0.05$) more nouns from a set emerged as identical to that observed in the discussion of gross *known* word sums i.e. a total of 48 hours:

Conclusion 39: The difference in the time over which target word presentation need occur for a statistical difference in noun totals to emerge must exceed or equal 48 hours.

The McNemar 'pairwise' test findings that so emphatically denied spaced learning as a cause of differences in *known* sums of verbs, adjectives and adverbs[97] from reading a set of a pair, agree with the vote-count/sign test when applied to these same lexical classes. For words in State 6, for example, the test finds no indication of a 'general' spaced learning benefit, irrespective of word type examined; see Section 5.11.1 for details (recall, however, the concerns regarding test sensitivity mentioned in footnote 89). For sums of words in State 5, the evidence for a 'general' spaced learning advantage emerges for nouns alone ($p=0.021$). For verbs (State 5) the test yielded a p-value of 0.344, for adverbs (after apportioning two ties) a p-value of 0.344, and for adjectives a p-value of 0.109. Nouns aside, these findings support the null hypothesis (Section 4.22.1) contention that spaced encounters with target words do not account for statistically significant ($p<0.05$) differences in gains of any particular word class. The significant differences in gross *known* word sums (i.e. the *known* total undifferentiated by class) for pairs of sets where such differences indeed arose (see Section 5.5.1) stems more from a spaced learning benefit raising the sum of *known* nouns than it does an impact, if

[97] The pairwise tests revealed no such cases.

any, upon sums of verbs, adjectives or adverbs. Yet only from 4 out of the 10 possible pairs of sets, however, did children gain significantly ($p<0.05$) more words in aggregate from reading a set of a pair, leaving a substantial six pairs from which they did not. Of these four cases, only from two, i.e. Sets 5 and 2 (i.e. case **b,** $_{6+5,\ 5\&2}$), and Sets 5 and 3 (i.e. case **c,** $_{6+5,\ 5\&3}$) did a statistically significant difference in extra *known* nouns contribute markedly to the gross total of additional *known* words from reading one set rather than the other (45% and 42%, respectively, of those gross sums in these two instances). The actual differences in sums of words, by class, that children gained from their reading experiences appear in Figure 5.27, below. A '-' indicates more words gained from the set offering the relatively massed target word presentation of the two.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1		**State 5** n-3 v0 adj1 adv2	**State 5** n1 v7 adj-5 adv4	**State 5** n2 v5 adj2 adv4	**State 5** n4 v5 adj5 adv2
Set 2			**State 5** n4 v7 adj-6 adv2	**State 5** n5 v5 adj1 adv2	**State 5** n7 v5 adj4 adv0
Set 3				**State 5** n1 v-2 adj7 adv0	**State 5** n3 v-2 adj10, adv-2
Set 4					**State 5** n2 v0 adj3 adv-2
Set 5					

Figure 5.27: Differences in the number of words, by class, in State 5 from reading alternative pairings of sets.

The binomial sign test findings (Section 5.8.3) of significantly ($p<0.05$) more State 5 words in aggregate (i.e. not differentiated by class) gained in cases (**a**,$_{6+5,\ 5\&1}$) and (**b**,$_{6+5,\ 5\&2}$) agrees with the vote-count/sign test results that likewise reveal typically larger State 5 word gains from spaced learning opportunities. Applied to the totality of data in Figure 5.27, and after apportioning zeros (see Section 4.21.1), the test returned a p-value of 0.000. The finding represents a strong affirmation of the sensitivity of State 5 sums to presentation time manipulation.

<u>Conclusion 40</u>: The sum of words in State 5 (undifferentiated by lexical class) is sensitive to the time over which children encounter those words in the experimental texts. This does not necessarily hold true for each lexical class considered in isolation.

5.12.4 Is there a general effect?

What does the vote-count/sign test reveal if applied to words of the same class occupying States 6 and 5, as opposed to sums occupying one state or the other (Section 5.12.5)? For nouns, and despite the vote-count/sign tests affirmation of a spaced learning effect upon State 5 word gains ($p<0.021$), the sums of N+ N- pairs miss significance by a wide margin ($p=0.343$). Explanations include: (1.) An insensitivity of State 6 noun totals to more, or less, spaced learning that conceals an effect on State 5 noun sums in the manner that State 4 words potentially do so when comparing total words learned (the sum of nouns, verbs, adjectives and adverbs) under the State 6+5+**4** and State 6+5 notions of *knowing* (see Section 5.9.1) or, (2.) That a spaced learning advantage operative upon State 6 nouns only emerges should additional target word presentation time exceeds 72 hours. If the latter, then the vote-count/sign test would appear insufficiently responsive to detect from the available data what nevertheless amounts to a real spaced learning advantage. If 'insensitivity' (possibility 1), conversely, seems the more plausible, then the spacing effect has little or no impact on children's learning outcomes should we choose to define *known* nouns exclusively as occupants of States 6+5. Which explanation most convincingly accounts of the vote-count/sign test's null hypothesis affirmation (i.e. no spacing effect if *known* nouns comprise occupants of States 6+5) remains unclear given the limited data available. That children gained more nouns to the State 6 standard from reading Set 5 rather than Set 4, and again from Set 4 as opposed to Set 3, suggests spaced learning may indeed contribute to noun gains in these cases despite statistical indications to the contrary (Section 5.11.1). This tentative suggestion, as we have seen, receives a certain modicum of support from the ever increasing addition to words familiar to the State 6 generally (i.e. the aggregate of those in States 6+5) with each 24-hour extension to presentation time beyond 72 hours (see Figure 5.1, p.148). Such a ‘pattern’ of additions beyond the ‘72 hour’ mark (Section 5.8.1) unfortunately represents a data profile to which sign tests, restricted to drawing upon simple comparisons of binary values as input, necessarily remain insensitive. A replicative study with a larger number of participants would likely provide some useful clarifications.

What of verbs? The participant group gained more verbs (to the 6+5 test of *knowing*) from the set of a pair providing the longer target word presentation time in 10 cases. The vote-count/sign test reveals a significant difference (p=0.002) in the sums of N+ and N- pairs (i.e. 10 and 0), despite the contrary McNemar ‘pairwise’ results affirming the participant group’s failure to gain significantly more of this class from any set of a possible pairing (Section 5.12.1, above). The findings indicate a ‘general’ spacing effect yet one apparently insufficiently robust to reveal itself from ‘pairwise’ comparisons.

As for adjectives among VSAT States 6+5, children gained numerically more words from the set of a pair offering the relatively massed learning opportunity in two cases (out of the possible 10), and from that providing the more spaced learning in the remaining 6 pairs. From the pair of Sets 1 and 2, the participant group gained an identical sum from reading each set, as indeed they did from reading Sets 4 and 5. After apportioning the two ‘zero difference’ cases equally among the N+ and N- observations, the *known* adjective sum now misses significance comfortably at p=0.344.

Regarding the fourth content word class, adverbs, the vote-count/sign test revealed children as a group learned additional words from the set offering the more distributed presentation in 7 cases, and identical sums in 3. Allocating one zero case to the N+ sum, and two to the N- total, albeit biasing the test towards null hypothesis preservation (Section 4.21.1), gives a score of eight N+ and two N- values. Applying the vote-count/sign test to these figures yields a p-value that comfortably misses statistical significance (p=0.109).

<u>Conclusion 41</u>: Even though ‘pairwise’ (McNemar) tests failed to yield evidence of a spacing effect operative at the level of word class (with *known* words as those in States 6+5), the vote-count/sign test suggests the likelihood of such

an effect operative on the class of verbs.

<u>Conclusion 42</u>: A spaced learning effect could operate generally on the class of verbs as indicated by the vote-count/sign test results, but not always sufficiently to reveal itself in set versus set, i.e. pairwise, comparisons of gains.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1		<u>6+5</u> n-5 v1 adj0 adv2	<u>6+5</u> n-3 v5 adj-2 adv4	<u>6+5</u> n0 v6 adj4 adv4	<u>6+5</u> n5 v7 adj4 adv4
Set 2	<u>6+5</u> n-5 v1 adj0 adv2		<u>6+5</u> n2 v4 adj-2 adv2	<u>6+5</u> n5 v5 adj4 adv2	<u>6+5</u> n10 v6 adj4 adv2
Set 3	<u>6+5</u> n-3 v5 adj2 dv4	<u>6+5</u> n v adj adv		<u>6+5</u> n3 v1 adj6 adv0	<u>6+5</u> n8 v2 adj6 adv0
Set 4	<u>6+5</u> n0 v6 adj4 adv4	<u>6+5</u> n2 v4 adj2 adv2	<u>6+5</u> n3 v1 adj6 adv0		<u>6+5</u> n5 v1 adj0 adv0
Set 5	<u>6+5</u> n5 v7 ad4j adv4	<u>6+5</u> n10 v6 ad4j dv2	<u>6+5</u> n8 v2 adj6 adv0	<u>6+5</u> n5 v1 adj0 adv0	

Figure 5.28: Number of words of each lexical class in State 6+5.

5.12.5 Known words by class (6+5), and measures of gain; Research Question 2

The study identified two cases –arguably three, if we include case (**a**, $_{6+5,\ 5\&1}$)– in which children learned significantly ($p<0.05$) more words of the same class from reading a set of a pair. These cases were: (a.) nouns from reading Sets 5 and 2, and (b.) nouns from reading Sets 5 and 3. A full summary of participants' gains appears in Figure 5.29, the blank cells representing combinations of sets unassociated with significant differences in *known* word totals. The disparity in the totals for learned nouns might seem impressive, ranging from ten in case (**b**, $_{6+5,\ 5\&2}$) to eight in (**c**, $_{6+5,\ 5\&3}$), while the difference in the average number of *known* words of this class ranges from 0.36 to 0.28. The percentage of nouns children learned from their reading appears 'large,' albeit variable. A referral to Table 5.2 in Section 5.5.1 indicates children as a group gained 68% of the total possible nouns from Set 3, as opposed to 96% from Set 5 (case **c**, $_{6+5,\ 5\&3}$), this amounting to 42% more nouns from the relatively spaced vocabulary presentation Set 5 provided. From reading Set 2 participants gained 60.7% of target nouns. The increase in nouns arising from the more spaced learning condition amounted to 59%.

<u>Conclusion 43</u>: In cases where children gained significantly more nouns from reading a set of a pair, the percentage increase in the sum of words of this class attributable to differences in the time intervals between word encounters ranged from 42% to 59%.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					
Set 2					Difference in the sum of nouns gained: 10 Difference in the average of noun gains: 0 .36 Percent increase from having read Set 5: 59%
Set 3					Difference in the sum of nouns gained: 8 Difference in the average of noun gains: 0.28 Percent increase from having read Set 5: 42%
Set 4					
Set 5					

Figure 5.29: Differences in the averages, *known* sums and percentages of target words gained (6+5), by class.

5.13 *Known* words as those which occupancy VSAT States 6+5+4 (Research Question 2)

For this, the most inclusive definition of *known* employed in the present study, Friedman tests (Section 5.6.1) and follow up binomial sign tests, revealed two cases of significant difference in *gross* word totals after apportioning zero scores:

1. case (**a,** $_{6+5+4,\ 5\&1}$): from reading Sets 5 and 1 (p=0.01)

and

2. case (**b,** $_{6+5+4,\ 5\&2}$): from reading Sets 5 and 2 (p=0.00)

5.14 Observations on cases

For case (**a**, $_{6+5+4,\ 5\&1}$) the McNemar ('pairwise') test returned a significant difference in the *known* sums of one class only, verbs (p=0.039), adjectives (p=1.00) and adverbs (p=0.48) each missing significance by comfortable margins, and nouns only narrowly (p= 0.063). In the second case, case (**b**, $_{6+5+4,\ 5\&2}$), McNemar (pairwise) testing indicated significant differences in the *known* word totals of two lexical classes: (1.) verbs (p=0.031), and (2.) nouns (p=0.016). For both adjectives and adverbs the test supplied identical p-values of 1.000, implying learning insensitivity to presentation time adjustments.

The participant groups' general profile of noun and verb gains, along with their distribution among States 6, 5 and 4 appears in Figures 5.30 and 5.31 for Sets 5 and 1, and 5 and 2, the sole pairings from which the participant group gained significantly more words (undifferentiated by class) from spaced target word presentation. The likely cause of the differences in nouns and verb totals in cases (**a,** 6+5+4, 5&1 and **b,** 6+5+4, 5&2) now becomes a little clearer, relative proportions (and sums) of *known* words occupying each of the VSAT states visually apparent.

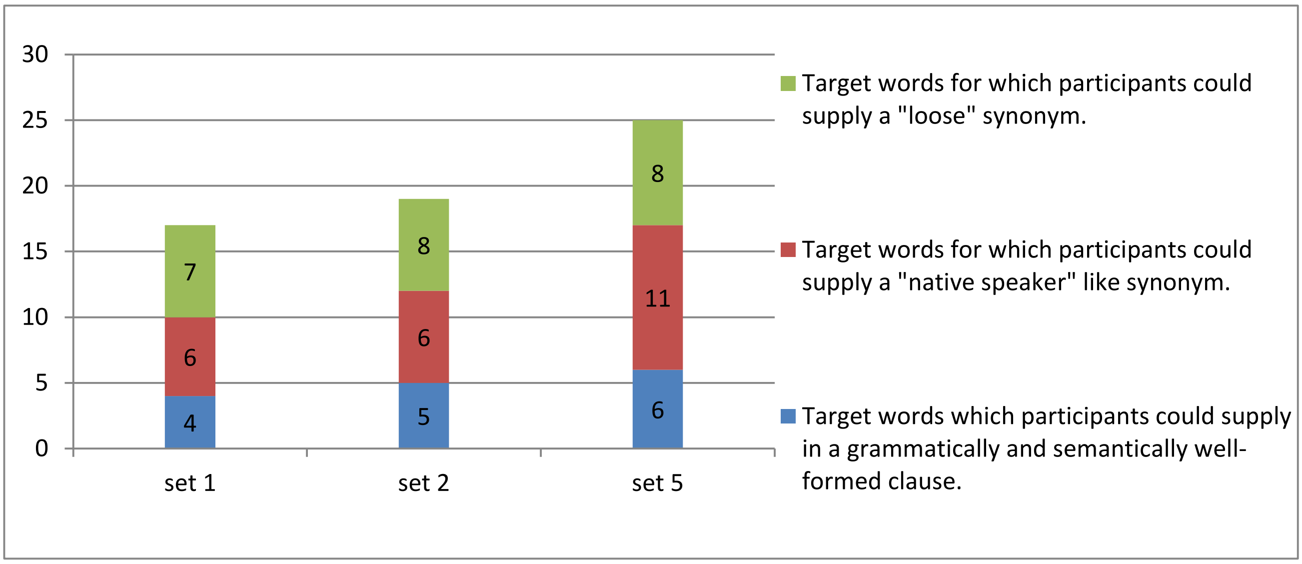


Figure 5.30: The proportion of verbs in States 6, 5 and 4 (cases **a**, 6+5+4, 5&1 and **b**, 6+5+4, 5&2).

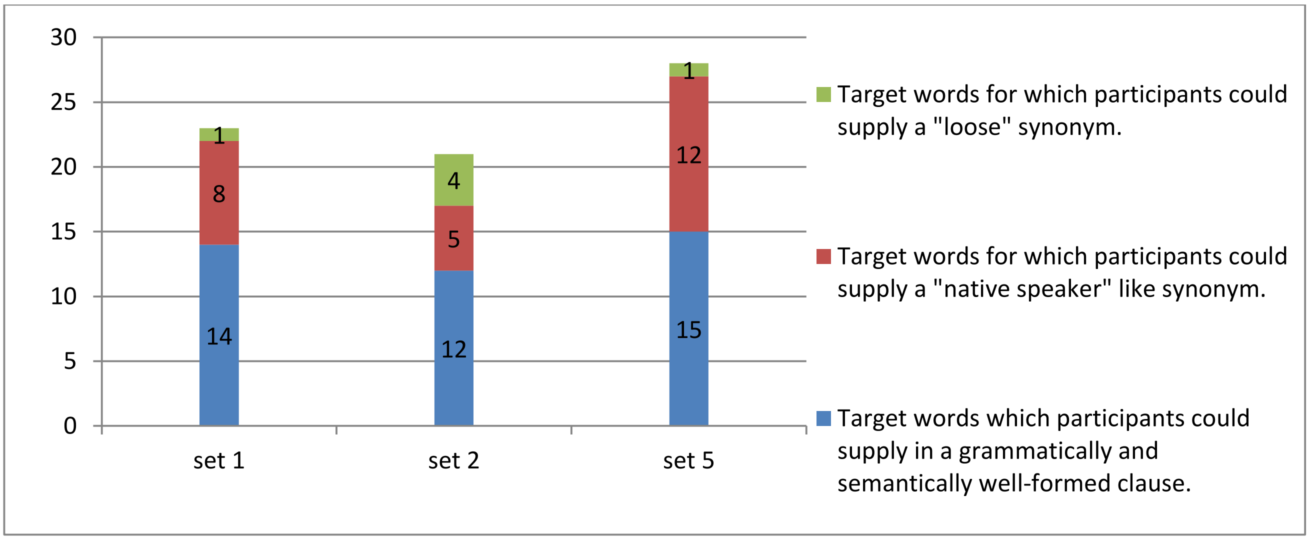


Figure 5.31: The proportion of nouns in States 6, 5 and 4 (cases **a**, 6+5+4, 5&1 and **b**, 6+5+4, 5&2).

How do the known totals 'break up' by lexical class? For verbs, sums of target words occupying State 6 from having read Sets 5, 1 and 2 miss significance comfortably, McNemar 'pairwise' tests returning p-values of 1.00 for case (**a**, $_{6+5+4,\ 5\&1}$) and p=0.581 for (**b**, $_{6+5+4,\ 5\&2}$), 'mirroring' the insensitivity of overall word totals (i.e. the aggregate of all four lexical classes) children *knew* to the State 6 standard (Section 5.8.1). A McNemar ('pairwise') test applied to *known* verb sums in State 5 adds further support to the null hypothesis (no spaced learning advantage), the test returning unremarkable p-values of 0.302 for the respective verb totals in Sets 5 and 1, and p=0.227 for Sets 5 and 2. For State 4, the test supplied an identical non-significant p-value of p=1.00 for verbs in cases (**a**, $_{6+5,\ 5\&1}$) and (**b**, $_{6+5+4,\ 5\&2}$) alike. Despite clear and unambiguous evidence (the aforementioned p=0.039 and p=0.031) for significant differences in *known* (the State 6+5+4 standard) verb totals in cases (**a**, $_{6+5+4,\ 5\&1}$) and (**b**, $_{6+5+4,\ 5\&2}$), when we look at the same VSAT state across each set of a pair (State 5 in Set 5 versus, State 5 in Set 1 etc.) children failed to gain statistically significantly more verbs from reading one set or the other.

That children collectively *knew* significantly ($p < 0.05$) more verbs overall (the total in States 6,5 and 4) from reading a set of texts of a pair, yet not significantly more occupying the same VSAT state 'across' the two sets, seems 'odd' and unintuitive. What, then, could account for such an apparently anomalous result? The obvious explanation may lie less in the absence of a spaced learning advantage operating the single VSAT state level, than simply the insensitivity of McNemar tests given so few words within each state in the respective sets to which the test was applied –that is, the non-significant outcome stems from the test's insensitivity given the limited data, this giving rise to a type 2 error (i.e. a false negative). The suggestion that a spaced learning advantage could indeed account for differences in *known* verb sums comes from results of the 'general' test procedure applied to the sums of N+ and N- pairs of sets from which children gained either more, or less, verbs to the 6+5+4 standard. The results imply a strong, pervasive, spaced learning advantage (p=0.002), perhaps extending across all VSAT States and potentially therefore accounting for additional verbs from a set providing the more spaced learning opportunity –indeed, the test supplied similar evidence for just such an effect when applied to verbs familiar to the States 6+5 test of *knowing* (p=0.002; see Section 5.5.3). In the light of this general (sign test) test finding, the absence of significant differences in *known* verb totals observed when comparing *known* verb sums in the same state across different sets of texts (from pairwise comparisons) may stand as no more than an artifact of the small sums of verbs serving as input to the McNemar ('pairwise') tests (see e.g. Hollander & Wolfe, 1999).

Conclusion 44: In those cases where the participant group gained statistically significantly more verbs from reading a set (of texts) of a pair, the source of that significant difference remains unestablished though plausibly lies in a general increase in verbs children *knew* to the standard of VSAT States 6, 5 and 4.

Nouns proved the only word class, aside from verbs, of which participants gained significantly (p<0.05) more from reading one set of a possible pair, the data showing that children gained more (p<0.05) nouns from Set 5 than Set 2 (case **b**, $_{6+5+4,\ 5\&2}$; p=0.016). For Sets 5 and 1, i.e. (case **a**, $_{6+5+4,\ 5\&1}$), the study finds that that respective noun sums miss significance narrowly (p=0.063) raising again the question of how reasonable becomes null hypothesis rejection given a broader, more holistic, view of the available evidence. The dilemma resolves into a question of probabilities: Does the 1 in 16 likelihood that the noun sums disparity arise by chance (implied by p=0.063), implying a valid null hypothesis, allow denying a spaced learning effect given the less than 1 in 20 probability (p<0.05) that conventionally suffices? Contesting the null hypothesis position is the significant difference the study identified in *known* noun sums from reading Sets 5 and 2 in case (**b**, $_{6+5+4,\ 5\&2}$). It seems surprising, arguably, that the 24-hour disparity in target word presentation associated with the participant group having read Sets 1 and 2 supplied apparently strong evidence for a spacing effect in case (**b**, $_{6+5+4,\ 5\&2}$; p=0.016), and yet not so in case (**a**, $_{6+5+4,\ 5\&1}$). If we agree this appears at least mildly paradoxical, then to claim a spacing effect accounts for differences in *known* noun sums in case (**a**, $_{6+5+4,\ 5\&1}$) becomes more tenable, the p-value (p=0.063) notwithstanding. This rests on two points: (1.) that the same Set, Set 5, provides the more spaced learning opportunity in cases (**a**, $_{6+5+4,\ 5\&1}$) and (**b**, $_{6+5+4,\ 5\&2}$) alike, and (2.) that the sets offering the more massed learning opportunities in these cases (Sets 1 and 2 respectively) differ only in that Set 2 provided 24 hours additional time (the smallest increment the study employs) over which target word presentation occurred. Children's additional noun gains in case (**a**, $_{6+5+4,\ 5\&1}$) would arise less obviously from spaced learning than those from having read Sets 5 and 2 (case **b**, $_{6+5+4,\ 5\&2}$), in this view, because the factors contributing towards differences in *known* noun sums (i.e. notably target word presentation time coupled with actual presentation time) depart more from the optimum in case (**a**, $_{6+5+4,\ 5\&1}$) than in case (**b**, $_{6+5+4,\ 5\&2}$). A spaced learning advantage, that is, indeed explains the extra nouns children gained from Set 5 in both cases, albeit contributing apparently less so to learning in case (**a**, $_{6+5+4,\ 5\&1}$). A test of this 'hypothesis,' albeit beyond the scope of the present study, would require exploring the apparent conduciveness of the massed learning Set 1 provides for learning nouns to the State 6 and 5 standard of *knowing* (hinted at in Figure 5.31) and its effect in compensating for the spaced learning advantage that Set 5 ensures for word learning generally. For the time being, it seems possible, though yet unproven, that such compensation explains the near miss of significance in *known* nouns in case (**a**, $_{6+5+4,\ 5\&1}$), the massed learning Set 1 provides apparently proving rather more favorable for noun learning than the less massed presentation (or more spaced) from children having read Set 2. The difference in State 6+5 nouns from reading Set 5 as opposed to Set 2 turns out quite substantial, amounting to 10 of the possible 28 nouns potentially available for the participant group to have learned. From reading Set 5 as opposed to Set 1, on the other hand, children gained an additional 5 nouns.

This suggestion that the participant group learned additional nouns from the more massed learning of Set 1, compared to Set 2 –in effect, that massed noun presentation proves more learning conducive than spaced– remains contentious, however, not least because pairwise comparisons of total words occupying State 6, whether of a particular lexical category, or gross sums undifferentiated by lexical class, conspicuously failed to identify significantly different learning outcomes from children having read any set of a possible pair (Section 5.4.3). The likelihood of relatively massed learning proving rather more learning supportive than spaced (or less massed) was noted in Section 5.8.1 which cited the case of fewer State 6 words (i.e. the sum of those in all four lexical classes) the participant group gained from reading Set 3 as opposed to Set 2, and the lesser sum still from reading Set 2 rather than Set 1. Only beyond 72 hours actual presentation time does a spaced learning advantage for words familiar to the State 6 standard conceivably emerge in the gross *known* word totals (see Section 5.4.1). The notion of Set 3 as a 'divisor' of sorts between sets of texts (namely Sets 1 and 2) from which children gained more from *reductions* in target word presentation time, and those sets (Sets 4 and 5) from which they gained more from presentation time additions seems at least 'visually' plausible from Figure 5.1 (p. 148). That noun totals proved not significantly different in case (**a**, $_{6+5+4,\ 5\&1}$), in this view, may arise from a spaced learning advantage associated with reading Set 5 providing insufficient gains to 'counterbalance' the relatively large number of nouns in State 6 from children's massed learning experience with Set 1. The source of the extra nouns from reading Set 5 in case (**a**, $_{6+5+4,\ 5\&1}$) remains, as under the under the States 6+5 definition of *known* for the same sets (5 and 1), the relatively large number of nouns which occupy VSAT State 5 (Figure 5.31) as opposed to States 6 or 4. The respective totals of nouns that occupy VSAT State 4 from reading Sets 5 and 1 (case **a**, $_{6+5+4,\ 5\&1}$) miss significance by a comfortable margin (p=0.58).

<u>Conclusion 45</u>: With *known* words defined as those in States 6+5+4, the effect of spaced learning resulted in a significant difference in the sums of nouns occupying VSAT State 5 from reading Set 5 as opposed to Set 2.

<u>Conclusion 46</u>: There are reasonable grounds to suppose that the extra nouns from reading Set 5 as opposed to Set 1 arise from a spaced learning advantage despite a p-value which exceeds the conventional value for alpha.

5.15 Distribution of known words among states

How substantially the additional words of a class (or classes) contribute to the extra words overall (undifferentiated by class) from the more 'learning conducive' set of a pair corresponds to the proportion of words of that particular 'class' (or classes) to the gross total. In case (**b,** $_{6+5+4,\ 5\&2}$), the 7 extra nouns and 6 verbs make up a not insubstantial 65% of the 20 additional words (the sum of all types) attributable to having read Set 5. For case (**a**, $_{6+5+4,\ 5\&1}$), children *knew* significantly (p<0.05) more words of only one type, verbs (p=0.039), the 8 extra words of this class contributing 35% of the

23 additional words the participant group gained from the Set 5 texts. The five extra nouns in this same case comprise 22% of the extra learned words, the six adjectives 26%, and the four adverbs, 18%. Acknowledging nouns as a lexical class from which participants gained significantly ($p<0.05$) more words from relatively spaced learning opportunities (‘Pairwise’ tests returned a p-value which misses significance only marginally, as we have seen), then the proportion of the total *known* words belonging to those classes responsive to a spacing effect benefit (for case **a**, $_{6+5+4,\ 5\&1}$) rises to a more impressive 56.5%. The extra nouns and verbs, the two classes from which children gained significantly more words from Set 5, now comfortably exceeds the combined sum of those constituting the remaining word types –adjectives and adverbs; words of these latter classes make up the remaining 43.5% of the extra words the participant group gained from their Set 5 reading experience.

In cases (**a** $_{6+5+4,\ 5\&1}$) and (**b**, $_{6+5+4,\ 5\&2}$), and notwithstanding the apparent insensitivity of learned adjective and adverb sums to spaced learning presentations, the possibility of a spacing effect contribution to the additional adjectives and adverbs from reading Set 5 as opposed to Sets 1 or 2, can’t quite as yet be discounted. The small number of study participants rules out any such forthright denial. Even statistically significant ($p<0.05$) differences in the *known* word sums of a lexical class, however, may not imply pedagogically meaningful additions over and above gains from the set of texts from which a child learned fewer (the issue of pedagogical implications is the subject of Part 3). For case **b**, $_{6+5+4,\ 5\&2,}$ for example, the additional nouns associated with reading Set 5 amounted only to an arguably meagre 7, or 0.25 extra per child attributable to spaced learning. Two further conclusions follow:

<u>Conclusion 47</u>: The primary source of the differences in totals of *known* words overall (undifferentiated by class) is a distributed learning effect which either (1.) expressed itself in a general increase in the sums of all four target word types, or (2.) disproportionately larger gains of verbs and/or nouns.

<u>Conclusion 48</u>: That spaced learning results in statistically significantly more gains of words of one lexical class from reading a set of a pair, does not imply statistically significant differences in the sums of words of another class from reading the same two sets of texts.

The additional nouns and/or verbs in States 6 and 5 that emerge in cases (**a**, $_{6+5+4,\ 5\&1}$) and (**b**,

$_{6+5+4,\ 5\&2}$) formed the subject of Section 5.12.1, leaving only words occupying State 4 to have therefore escaped comment thus far. With so few words of the same class occupying the latter state, ‘pairwise’ (McNemar) tests failed to reveal significant differences in the totals for any of the four word types examined. The finding affirms the null hypothesis (i.e. no spacing effect) position while not quite ruling out a possible effect had the study involved rather more than the 28 participants. The vote-count/sign test on State 4 sums likewise failed to identify a spaced learning advantage, with p-values missing significance by a wide margin for all four lexical classes: nouns, p=1.00; verbs, p=1.00; adjectives 0.754; and adverbs, p=1.00. Indeed, in several instances we see that the participant group gained an equal or additional sum of words *known* words to the State 4 standard from the more massed learning condition a set provided (see Figure 5.32). From Set 2, for example, children gained both more nouns and adjectives than from Set 4, and more nouns from reading Set 3 than from Set 5. Even the largest difference in *known* word sums to arise from the set of a pair offering the relatively spaced learning opportunity proved arguably quite ‘small,’ however, at just 5 extra words; this compares to a maximum gain from the set providing the more massed presentation of 6. The non-responsiveness of State 4 sums to presentation time leaves the sensitivity of State 5 sums, and possibly State 6, as the primary contributors to statistically significant differences in *known* totals of any class where such differences arose from McNemar ‘pairwise’ testing.

The apparent non-reactiveness of State 4 sums to presentation time manipulation reveals itself again when applying vote-count/sign test to the totality of data in Figure 5.32. The cases in which the participant group gained additional words from the set providing the more spaced learning amount to 16, and those for which relatively massed learning saw the higher gains, 17. After apportioning 7 zero difference pairs, the sum of N+ cases equals 19, and N- cases 21. Inputted to the vote-count/sign test, this supplies a p-value of 1.00. From the practical perspective of classroom teachers, the finding highlights once again a disproportionate contribution of spaced learning opportunities depending upon both lexical class and one’s particular sense of what ‘*word knowing*’ might reasonably imply.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1		**4** n2 v1 adj5 adv-4	**4** n4 v0 adj5 adv-1	**4** n1 v-1 adj-1 adv-3	**4** n0 v1 adj2 adv0
Set 2			**4** n2 v-1 adj0 adv3	**4** n-1 v0 adj-6 adv1	**4** n-2 v0 adj-3 adv4
Set 3				**4** n-3 v-1 adj-6 adv-2	**4** n-2 v1 adj-3 adv1
Set 4					**4** n-1 v0 adj3 adv3
Set 5					

Figure 5.32: Number of words of each lexical class in State 4 (Cases in which children gained more under more massed learning are prefixed by a minus sign).

5.16 Is there a general effect?

One issue remains –the likelihood that a pervasive spaced learning effect accounts for the differences in *known* sums by class (i.e. of those familiar to the States 6+5+4 standard) that eludes detection with McNemar (‘pairwise’) testing. For verbs, the vote-count/sign test strongly implies a pervasive effect, returning a p-value of p=0.002, despite ‘pairwise’ comparisons indicating significant p-values in just two instances, i.e. from reading Sets 5 and 1, and 5 and 2 (i.e. case **a**,$_{6+5+4,\ 5\&1}$; and **b**,$_{6+5+4,\ 5\&2}$). For nouns, no ‘general’ effect arises (p=0.109), albeit McNemar (‘pairwise’) tests reveal the participant group gained statistically more words of this class from Set 5 in case (**b**, $_{6+5+4,\ 5\&2}$). Nor did the vote-count/sign test suggest a spacing effect that might explain differences in adjective or adverb totals, the test returning respective p-values of p=0.754 and 0.109. This leaves two word types, then, nouns and verbs, as the sole classes for which one or other of the two tests the study employs (‘pairwise’ or ‘general’) reveals a spaced learning contribution to differences in learned word sums from having read a set of a pair. For the remaining lexical classes the case remains unestablished.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1		**6+5+4** n-2 v2 adj5 adv-2	**6+5+4** n1 v5 adj3 adv3	**6+5+4** n1 v6 adj3 adv1	**6+5+4** n5 v8 adj6 adv4
Set 2			**6+5+4** n3 v3 adj-2 adv5	**6+5+4** n3 v4 adj-2 adv3	**6+5+4** n7 v6 adj1 adv6
Set 3				**6+5+4** n0 v1 adj0 adv-2	**6+5+4** n4 v3 adj3 adv1
Set 4					**6+5+4** n4 v2 adj3 adv3
Set 5					

Figure 5.33: Number of words of each lexical class in States 6+5+4.

5.16.1 Known words by class (6+5+4), and measures of word gain; Research Question 2.

Figure 5.34 reports gains from distributed learning expressed as sums, averages and percentages for those pairs of sets from which children gained significantly ($p<0.05$) more words from one or the other. From reading Set 5 and Set 2 (**b**, $_{6+5+4,\ 5\&2}$), for example, participants gained an average of 0.25 extra nouns and 0.21 additional verbs from the Set 5 reading experience. In case (**a**, $_{6+5+4,\ 5\ \&1}$) gains amounted to 8 additional verbs from Set 5, or an average of 0.29 extra words (per child) over and above those from Set 1. The proportion of verbs the participant group gained from having read Set 5 and Set 1 (of the maximum 28 gainable) indeed corresponded to an arguably impressive 89% and 60% respectively. But what, then, do the data reveal of observing a spaced learning advantage? The 'missing' cells in Figure 5.34 helpfully establish from just how few pairs of sets children gained statistically (p,0.05) more words during their RR sessions. For nouns, of the 10 possible cells (those to the right of the shaded diagonal), only 1 (or 10%) depicts a pairing of sets from which children learned additional words of this class from relatively more spaced learning opportunities. For verbs, the figure amounts to 2 cells, or just 20% of the total possibilities while for the remaining classes the percentage stands at zero. As a rule, the participant group rarely gained significantly more words of any class from reading a set of a pair, only doing so if the target words constituted nouns or verbs and if the pairs of sets consisted of combinations of Sets 1, 2 and 5. Of the 40 possible pairwise comparisons of *known* word sums by class (10 each for nouns, verbs, adjectives and adverbs), in just 3 instances (i.e. 7.5%) did McNemar tests reveal a significant difference in totals, or four (10% of the total possibilities) if one accepts a spacing effect contribution to the additional nouns in the controversial case (**a**, $_{6+5+4,\ 5\ \&1}$). For the most part, then, the likelihood of the participant group gaining significantly more words of a class from reading one set as opposed to another would appear low. In those cases where we observe a statistical difference at all, this is only for 2 of the four lexical classes examined.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					Difference in the sum of verbs: 8 Difference in average of verb gains: 0.29 Percentage increase, verbs: 47
Set 2					Difference in sum of verbs: 6 Difference in average of verb gains: 0.21 Percentage increase, verbs: 31 Difference in sum of nouns: 7 Difference in average of noun gains: 0.25 Percentage increase, nouns: 33

Figure 5.34: Differences in the averages, *known* sums and percentages of target words gained (6+5+4), by class.

5.17 General conclusions on differences in sums of learned words from reading the experimental sets.

Figure 5.35 condenses into a single display the learning outcomes recorded under all three definitions of *known* word (6, 6+5 and 6+5+4) to depict the totality of circumstances under which children gained more words of a class for the three definitions of interest. The conditions associated with a spaced learning advantage now become a little clearer, revealed by the sheer absence of simple ‘patterns’ or ‘consistencies’ in the display. We see from reading Sets 5 and 3, for example, that children gained statistically significantly more nouns from Set 5 (‘pairwise’ tests, p=0.008) under the States 6+5 test of *knowing* (Section 5.12.1) and yet failed to do so when defining *known* more broadly as those occupying States 6+5+4 (p=0.125); likewise, participants gained significantly more verbs from Set 5 than they did from reading Set 1, and yet not more nouns; the participant group learned more verbs from the 48-hour difference in target word presentation time associated with reading Sets 5 and 3 while failing to do so from reading either set of the pair 1 and 3. Firm, generally applicable conclusions, however, remain elusive. That children gained the larger sum of *known* words from Set 5 would seem a ‘constant’ of sorts, as does the absence of a spacing effect operative upon sums of adverbs and adjectives. The insensitivity of State 6 sums to spacing also stands as another general finding as does the broad range in presentation time differences (anything between 48 and 96 hours) associated with significantly more learning. Overall, however, Figure 5.35 falls well short of identifying exceptionless propositions. Whatever the role of additional presentation time on learning gains may amount to, little understood moderating factors impact upon spaced learning efficacy. Two such factors, *actual* presentation time and *difference* in presentation time, received attention earlier (see p.199, for example); the identities of other factors, their number, and their contribution remains unknown. With so few truly concrete findings to draw upon, the data on VSAT elicited word gains leaves ample scope for argument. Depending upon the test of *known* word one applies, the same difference in word presentation time may, or may not, explain significant differences in noun or verb totals. Nor, it seems, do gains appear so impressive that teachers would necessarily agree upon how pedagogically useful is the additional vocabulary from those sets of texts more learning conducive than others. To some, the total of additional words from reading Set ‘x’ as opposed to set ‘y’’ may appear ‘small,’ to others, ‘moderate,’ and to yet others still, possibly ‘large.’

	Set 5
Set 1	**States 6+5+4** Difference in the sum of verbs: 8 Difference in the average of verb gains from reading the sets of a pair: 0.29 Increase in verbs gained: 47%
Set 2	**States 6+5** Difference in the sum of nouns: 10 Difference in the average of noun gains from reading the sets of a pair: 0.36 Increase in nouns gained: 58% **States 6+5+4** Difference in sum of verbs: 6 Difference in the average of verb gains from reading the sets of a pair: 0.21 extra words from more spaced learning. Increase in verbs gained: 31% **States 6+5+4** Difference in sum of nouns: 7 Difference in the average of noun gains from reading the sets of a pair: 0.25 extra words from more spaced learning. Increase in nouns gained: 33%
Set3	**States 6+5** Difference in the sum of nouns: 8 Difference in the average of noun gains from reading the sets of a pair: 0.28 extra words from more spaced learning. Increase in nouns gained: 42.1%

Figure 5.35: Combinations of sets associated with a significant difference in '*known*' nouns, verbs, adjectives or adverbs.

Other observations from figure 5.35 help define the pedagogical gains spaced learning might offer should teachers embark on adapting reading materials to the specifications of one or another set of those the current study employs:

<u>Conclusion 49</u>: For any definition of *known* word, the difference in the average number of words of a *class* for any pair of sets associated with significantly more learning, was less than 1 word (from the total of four encountered); the range of values for these differences lay between 0.21 (verbs from reading Sets 5 and 2, *known* words as those in State 6+5+4) and 0.36 (from reading Sets 2 and 5, with *known* words as those in State 6+5).

<u>Conclusion 50</u>: Of the 120 possible comparisons of *known* word sums from which a significant difference in sums of words of a single class could potentially have been observed (i.e. 30 possibilities given the State 6 test of knowing, 30 under the States 6+5 test and 30 under the States 6+5+4 test) in only five cases did a significant difference emerge.

<u>Conclusion 51</u>: Irrespective of the definition of *known,* a spaced learning advantage only accounted for a significant difference in the sums of *known* words of two lexical classes –these classes were nouns and verbs.

<u>Conclusion 52</u>: The maximum percentage increase in *known* word totals (of a particular word class) from reading the set of any pair associated with the more substantial word gains was 58%. The minimum was 29%.

Issues remaining unresolved include (1.) whether the notion of *known* affects how substantially the word gains of a particular word class contribute to the additional words *overall* (the aggregate of those from all four lexical classes) arising from more spaced learning opportunities,[98] and (2.) The responsiveness of *known* sums of each word class under alternative tests of *knowing,* should target word presentation time exceed the 96 hour maximum the study examines. That the vote-count/sign test findings supply no indication of State 4 word sums responding to word presentation time manipulation (see p. 215) raises the obvious question of just why this is so. Among the similarly unresolved questions in regard to VSAT States 6 and 5 remain the effects upon learning outcomes from 24-hour increments to presentation time beyond five days and, in particular, the point at which children might begin to gain statistically significant ($p<0.05$) additional words of a class from more massed learning opportunities. The possibility of rather more State 6 word gains from additions to

[98] That is, the ratio of words of a particular class to the sum of extra words overall a reader gained from the set from which s/he learnt most.

presentation time exceeding 96 hours was noted in Section 5.8.1, along with the likelihood of lesser additions to State 5 sums. Evidence for the latter appears in Figure 5.27 as the several instances in which children gained fewer state 5 words from that set of a pair exposing the reader to the more massed target word presentation. Research into relative learnability i.e. differences in the responsiveness of word totals in each VSAT State to presentation time manipulation recommends itself for further study, partly because it potentially yields altogether finer grained projections of word gains over months, academic terms or years, but also because breaking down of gains in terms of lexical competencies allows for objectively weighing the costs and benefits of textual adaptation. As we shall see (Part 3), even small differences in the likelihood of word gain for any lexical class or classes may express themselves in substantial effects on learning outcomes over the long-term (e.g. an academic year). Even so, the value of those gains depends upon individual notions of what *word knowing* properly implies.

<u>Conclusion 53</u>: A statistically significant difference in the sum of *known* words of a particular class could emerge (albeit initially absent) either from expanding the definition of *known* words such that the definition is more inclusive (i.e. less demanding) or, alternatively, from redefining the definition of *known* words such that it is less inclusive (i.e. more demanding).

5.18 Conclusion of Parts 1 and 2

Figures 5.36 amounts to a composite of Figures 5.25 and 5.35 and depicts all combinations of sets of texts from which children gained either (a.) significantly more target words overall (i.e. words undifferentiated by class), and/or (b.) more words of a specific lexical category. This overall summary of the conclusions for each Research Question serves two purposes: First, it reveals details not readily apparent from the findings applicable to each Research Question when viewed independently of the other. Second, it identifies the commonality in the circumstances under which children gained more words of a class or of *known* words overall i.e. it identifies conclusions which extend beyond any single notion of word *knowing*, to all.

1

	Set 1	Set 2	Set 3	Set 4	Set 5	
Set 1					**(6+5)** Sum:20 Av:0.72 Inc,:39.2%	**(6+5+4)** Sum:23 Av: 0.83 Inc.:33 **(6+5+4)** Verbs Sum: 8 Av: 0.29 Inc.: 47
Set 2				**(6+5)** Sum:16 Av:0.58 Inc.:32.6%	**(6+5)** Sum: 22 Av:0.79 Inc.:44.8 **(6+5)** Nouns Sum: 10 Av: 0.36 Inc.58.1%	**(6+5+4)** Sum: 20 Av:0.72 Inc.:33 **(6+5+4)** Nouns Sum: 7 Av: 0.25 Inc.:33% **(6+5+4)** Verbs Sum: 6 Av: 0.21 Inc.: 31%
Set 3					**(6+5)** Sum: 16 Av:0.58 Inc.:29% **(6+5)** Nouns Sum: 8 Av: 0.29 Inc.: 42.1%	

Figure 5.36: *Known* words. (*Sum=additional words from text offering the more distributed word presentation; Av= difference in the average number of words learned from reading the sets of a pair; Inc. = percentage increase in known words attributed to a spacing effect.).*

What does this summary reveal? Perhaps least controversially, we see that significant differences in gross totals of *known* words do not necessarily imply differences in gains of any one lexical class. With *known* words as those occupying VSAT States 6 or 5 the participant group failed to learn statistically ($p<0.05$) more nouns, verbs, adjectives or adverbs from Set 4 than Set 2 despite learning significantly more words overall (the sum of words of all four types); this case was singled out for discussion in Section 5.8.3. Readily evident, too, is children's (collective) failure to gain significantly more words (*known* = occupancy of States 6+5) of any lexical class or classes from reading Set 5 as opposed to Set 1, despite collectively having learned a significantly larger sum of words in total (i.e. the aggregate of nouns, verbs, adjectives and adverbs) from the Set 5 reading experience (see Section 5.8.2). Most importantly, perhaps, Figure 5.36 brings out the striking rarity of a spaced learning advantage in general, this clearly apparent from the many blank cells as a proportion of those containing data. Only from reading Sets 5 and 1, 5 and 2, 5 and 3, and 4 and 2 did children gain statistically significantly more target words (undifferentiated by class) from reading a set of a possible pair –an unimpressive 4 pairs (or 20%) from a possible 20[99]. From 16 pairs of sets (i.e. 80%), therefore, the participant group failed to gain significantly more words from relatively distributed target word presentations. If we now ass on the 10 pairs of sets from which children failed to learn additional ($p<0.05$) target vocabulary to the State 6 standard (Section 5.4.1) to the aforementioned 16, then the pairs of sets associated with significantly more gross learning (i.e. the total of *known* words regardless of lexical class) under any definition of *known* amounts to just 4 out of 30 (16+10), or 13% of the potential possibilities. Cases in which children gained significantly more words of a single lexical class from the set offering the relatively spaced learning opportunity of a pair, likewise proved 'few' –a mere 5 cases, the sum appearing less impressive still as a proportion of the total (120) from which they might have potentially made a significantly ($p<0.05$) larger gain. Only from 4% of the possible pairs of sets (i.e. 5/120), then, did the participant group gain significantly more words of a particular class from the set of a pair providing the more distributed target word presentation.

For both the aggregate of *known* word sums, and those by class, the study identifies instances where participants gained significantly ($p=0.05$) more target vocabulary under one notion of *knowing* ('pairwise' tests) while not another. The examples include the significant difference in sums of learned verbs associated with reading Sets 5 and 2 with *known* word defined as occupants of States 6+5+4, and the notable absence of any such statistically significant difference if restricting *known* words to those in either VSAT State 6 or States 6+5. An illustration of the same point, but for gross (undifferentiated) sums, comes from the participant group having gained significantly more words from Set 4 than from Set 2 when *known* denotes occupancy of States 6+5, and yet failure to do so with

[99] This consists of 10 pairs of sets given the 6+5 definition of *known* word, and 10 pairs under the 6+5+4 based definition).

known words defined as those in either State 6, or States 6+5+4.

The absence in Figure 5.36 of significant differences in sums of words *known* to the State 6 standard (though see Section 5.8.1) points to an insensitivity of these sums to presentation time manipulation that seems to hold as robustly at the word class level as it does to gross *known* totals. This leaves us with a general finding that State 5 sum sensitivity serves as the primary source of significant differences in word gains from reading a set of a pair should such differences emerge at all. Because the sum of words occupying any specific VSAT State could, however, prove more, or less, responsive to presentation time changes than the sum of those occupying another, an increase (or reduction) in the time over which one set of texts of a pair presented target words might impact differently upon learning outcomes depending upon the definition of *known* we acknowledge (1. State 6; 2. 6+5; or 3. 6+5+4). At the one extreme we have the example of words in VSAT State 6. The study finds no one set of texts emerged as more effective for developing the State 6 productive skills relevant to speaking and writing than did any other. The evidence comes from results of 'pairwise' and 'general' test procedures, albeit subject to caveats noted in Sections 5.4.2 and 5.11.1. Similar unresponsiveness exists for words in State 4, with both pairwise testing (McNemar tests) and the vote-count/sign procedures supplying p-values that strongly agree with the null hypothesis denial of a spaced learning advantage. On the other hand, sums of word occupying State 5 indeed could, and occasionally did, vary quite markedly depending upon an interplay of presentation time difference, actual presentation time, and the particular notion of *knowing* under review.

Just how spaced the encounters with target words need be before statistically significant differences in learning arise now becomes a little clearer. The study finds children did not gain significantly more words (either the gross sum of *known* words, or totals for words of the same lexical class) unless a set presented its target vocabulary over at least two additional days than did the other. That a 48-hour extension in target word presentation time falls well short of assuring a significant difference in word gains becomes readily apparent, however, from children having gained significantly more words in aggregate (the sum of nouns, verbs, adjectives and adverbs) from reading Set 4 as opposed to Set 2 under the 6+5 notion of *known*, and yet failing to do so from having read the pair of sets 1 and 4. In both these cases, the higher ordered set of each pair presented its target words over an additional 48 hours. A three-day (i.e. 72 hour) addition to presentation time saw children gaining significantly ($p<0.05$) more *known* words (in aggregate) from several set pairings under both the 6+5+4 and 6+5 tests of *knowing* but, also, significantly more learning of both nouns and verbs. A four day (i.e. 96 hour) addition proved sufficient for children to gain significantly more words overall from the more distributed learning opportunity (Set 5, as opposed to Set 1), but a significant difference in learned words of just one lexical category –verbs (known words as those occupying State 6+5+4).

<u>Conclusion 54</u>: A significant difference in word gains, whether in aggregate or of any single class, requires that target words be distributed over a minimum of two days more than that afforded by the set providing the more massed encounters.

Figure 5.36 provides fertile ground for argument. For example, one can readily see how gains from spaced learning might seem less than impressive from a practicing educator's perspective than to those merely intent on demonstrating an effect as such. Under both the 6+5 and 6+5+4 definitions of *knowing*, the maximum sum of additional words from spaced learning amounts to 22 and 23 respectively, corresponding to an increase of 33% (States 6+5), or 39.2% (States 6+5+4), more words than gained from relatively massed target word presentation. What, then, might such gains really mean to the language arts instructor? The minimum addition of *known* words from reading any pair of sets proves arguably disappointing at just 16, this with *known* words as those occupying States 6+5 and from children having read the pair of Sets 5 and 3, and 4 and 2; the increase amounts to a 42% word gain over those from the more massed learning in the former case, and a 32% gain in the latter. Compared to vocabulary gains from explicit instruction, textual adaptation to exploit spaced learning over the short term might strike teachers as offering rather little, especially so given the challenges textual adaptation involves (though see Part 3, below). The maximum difference in the sum of words of a single class children collectively achieved stood at just 10 (i.e. nouns from reading Set 5 as opposed to Set 2), albeit corresponding to an impressive 58% word increase over those from the massed learning opportunity Set 2 provided, and an average of 0.36 extra words per child.

Part 3

5.19 Research Question 1: Discussion of pedagogical significance

Part 2 examined significance from a statistical perspective that explored whether outcomes from an experimental condition arise other than from chance occurrences. The various commentaries on study findings summarized learning gains by referencing typical measures of educational attainment including gross sums, averages and percentages of words children *knew* to the standard of various VSAT descriptors. The current section focuses rather more on pedagogical importance from a practicing teacher's perspective, an issue barely touched upon, and then only peripherally, in the previous discussions. Drawing upon the results from Parts 1 and 2, the study estimates differences in learning over the course of one academic year assuming children were to read RR materials exclusively designed to the specifications of those sets of texts employed in the current investigation. To express findings in meaningful terms, the analysis draws upon the 'standard measure' (SM) construct introduced in Chapter 4, where a 'standard measure' refers to the sum of words children (objectively comparable to those who participated in the current study) might reasonably gain from unadapted RR texts during a single academic year. The methodology for determining SMs, and then comparing these to estimated learning gains from adapted texts involves four steps, as noted in Section 4.21.1, namely: **Step 1**. Identifying the sum of novel words a child encounters during RR sessions over one academic year (be they words of a particular class or aggregate totals); **Step 2**. Estimating the sum of words a child would likely gain to a criterion of *knowing* from reading sets of texts designed to the specification of those the current study employs (Sets 1-5); **Step 3.** Estimating the proportion of novel words (or differentiated by lexical class, if exploring Research Question 2) a child might gain from one year of

RR sessions from unadapted texts; and **Step 4**. Comparing the sum of words from steps 2 and 3 to reveal additional gains from reading a set of a pair in SM form.

5.19.1 Standard Measures and significance of word gains undifferentiated by word class (Research Question1).

_**Step 1**. *Identifying the sum of novel words a child is likely to encounter during RR sessions from unadapted (regular) texts over the course of one academic year.*

To estimate yearly gains calls for three statistics: (1.) the number of days over which RR takes place during an academic year, (2.) the duration of RR sessions, and (3.) the average of children's reading rate in words per minute. Assuming a typical 185-day academic year (not unusual among schools following the National Curriculum of England and Wales), and 4 days (from 5) of RR per week, then children will engage in RR for a total of 148 school days (i.e. 80% of 185). The time devoted to RR sessions will vary according to school policies, with 25 minutes suggested here as a reasonable 'ballpark' figure based upon conversations with other international school colleagues (in the host institution the figure was 35 with some time 'lost' to children selecting and/or returning books). Participants in the current study read at approximated to 147 words per minute (wpm), a rate comparable to that Carver (1983) noted of grade 4 students in American schooling, and similar to that the present researcher has recorded for English EAL nine-year-olds attending private schooling. The product of reading rate (147 wpm) and time spent on RR (i.e. 148 x 25 minutes) supplies the average child's reading volume. Given the above assumptions, volume therefore amounts to 543,900 words per academic year (i.e. 147x25x148), a figure similar to Fielding, Wilson and Anderson's (1989) estimate for the 65th percentile American fifth grader. Assuming literacy teachers will, in general, provide texts with the optimal sum of unknown words for vocabulary learning –around one or two per 100, according to Nation (2000)– then the typical EAL child encounters between 5,439-10,878 novel words per annum, or an average of 8,158. The latter figure will serve as the presumed total in the computations that follow.

Step 2. *Estimating the sum of novel words children go on to gain of the total such words encountered per year.*

In an SM analysis, word gains come across the more impressive to the extent they exceed the presumed annual total of words children gain from unadapted texts during RR sessions, leaving the researcher to determine this, the critical SM, from the evidence available. Following Nagy et al. (1985, 1987), the sum (the SM) roughly approximates to the product of the total unfamiliar words (UW) a child encounters per annum and the probability (P) of learning a word from a single textual meeting (i.e. annual sum = UW x P). The unfamiliar word total (UW), as noted, reasonably amounts to the 8,158 words computed above. For the relevant probability estimate (P), candidate figures appear in both Nagy et al. (1985, 1987) and in Swanborn and de Glopper (1999), values ranging from around 0.05 to 0.15. The discrepancies arise from differences in grade level, whether pretesting took place prior to reading sessions, and the text genre of the scripts the investigations employed (Nagy et al., 1987). Which among the available figures amounts to the most reasonable logically hinges upon just how similar the student population from which the probability emerges, to the particular population of children participating in the current study. Applying this, the 'similarity test,' the study adopts the (P) =0.085, appearing in Nagy et al. (1987), a figure derived from students and reading circumstances broadly comparable to those in the institution where the present investigation took place. Specifically, four points support this choice. First, the estimate originates in research with schoolchildren of comparable age to those in the present study –American grade 3 children (unfortunately, Nagy et al. (1987) neglect to mention the number and proportion of non-English L1 native speakers). Second, the figure assumes prior reader familiarity with the concept a novel word denotes –the same familiarity that participants in the present study could claim of target words in the experimental texts (see Section 4.14). Third, the estimate derives from exposure exclusively to narratives, the genre of the experimental texts in the current study. Fourth, the Nagy et al. (1987) notion of *known* seems to correspond broadly with word occupancy in VSAT States 6+5 or States 6+5+4; if indeed so, then *known* word sums would likely prove comparable whether derived from the VSAT or the test Nagy et al. (1987) employed. Whether the 6+5 or 6+5+4 notion of *known* corresponds most closely to the Nagy et al. (1987) criteria for assigning words a *known* designation depends upon how similar the lexical understanding sufficient to score ‘correct’ responses in Nagy et al.’s (1987) MCT based study, to the particular understanding that suffices to warrant a *known* ‘score’ from VSAT testing. On this, direct evidence remains lacking. A sense of the Nagy et al. (1987) conception of *known,* however, does emerge from descriptions the authors provide of their MCT instrument: for example, that distracters are not “meant to be tricky or extremely difficult” (p. 248), being comprised of “definitions … semantically similar to the target word and of the same part of speech.” This would imply a less demanding test of *known* than suffices to supply a word in a semantically and grammatically correct sentence (the requirement for VSAT State 6 occupancy) but, arguably, a competence corresponding to occupancy in VSAT States 6+5. Without further details to go on, scope for disagreement unfortunately remains. What exactly does “similar” denote in the context in which Nagy et al. (1987) employ the term? What does it mean to be “tricky” or “extremely difficult?” How different (and in what regard) need a word be before the label “similar” ceases to apply? Not least, it is unclear just how one might approach measuring ‘degree of similarity’ in objective terms. Given the uncertainties, the current study adopts the position that word *knowing,* as Nagy et al. (1987) conceive the term, corresponds most accurately to neither the 6+5 or 6+5+4 based definitions, but rather a synthesis of the two –a middle ground representing neither one nor the other.

The above in mind, working with the 0.085 probability and assuming children encounter 8,158 novel words per year, the sought-after standard measure (i.e. the product of 8,158 X 0.085) therefore amounts to 693 words. It is this figure that will serve as the presumed yearly estimation of gains from one year of RR from unadapted texts.

Step 3. *Estimating the proportion of novel words a child might gain from one year of RR sessions having read adapted texts during all RR sessions.*

Identifying differences in sums of novel words children *know* from reading each experimental set over one academic year begins with the probability of learning a novel word from a single encounter during a reading experience (see Nagy et al. 1985, 1987). To compute this from probabilities derived from multiple encounters with the same word (during the current investigation children encountered each target word on 12 occasions) the study applies the following relationship: $P_n = 1 - (1-P_1)^n$ where term Pn denotes the probability of learning from context given 'n' exposures (12 in this case), and p_1 the probability from just one meeting.[100] The p_1 estimations this relationship implies appear in Table 5.4 (below). The likelihood a child gains an unknown word from reading Set 1 type texts to the VSAT 6+5 standard of *knowing,* for example, amounts to 0.048. Were a child to have read Set 5 type texts the figure proves somewhat higher standing at marginally less than 8 percent (0.079). The highest probability of learning from a single meeting stands at just over 13%, this assuming the 'expansive' State 6+5+4 notion of *known,* and exposure to Set 5 type texts. The lowest probability, conversely, comes to 0.047 (marginally less than 5%) and derives from reading Set 2 type texts with *known* words restricted to those occupying VSAT States 6+5. Blank cells in Table 5.4 designate a set of texts that, for the notion of *knowing* of interest, (left column) children failed to gain significantly more, or less, target words from reading one Set than another.

	Set 1	Set 2	Set 3	Set 4	Set 5
6					
6+5	0.048	0.047	0.054	0.064	0.079
6+5+4	0.074	0.081			0.133

Table 5.4: Probability of learning a word from a single encounter under alternative definitions of known.

Table 5.5 displays the projected yearly totals based on the probabilities of learning from a single encounter (Table 5.4) assuming (1.) 8,158 novel word encounters per year (p. 287) and, (2.) that

[100] Informally, $(P-1)^{12}$ is the probability of an event not occurring on 12 out of 12 occasions. Therefore, $1-(p-1)^{12}$ is the likelihood of one success during those 12 occasions.

children received exposure to RR texts exclusively conforming to the specifications of a particular set the current study employs. From Set 1 type reading materials, for example, a child's predicted gain amounts to 391 words assuming we define *known* as word occupancy of VSAT States 6 or 5 (i.e. 0.048 x 8,158), and a more substantial 603 words under the more inclusive States 6+5+4 definition. From reading Set 5 type texts, the *known* sums prove somewhat higher with gains lying between 644 (States 6+5) and 1,085 words (States 6+5+4 test). Again, blank cells correspond to row (i.e. notions of word *knowing*) and column (sets of texts) intersections unassociated with statistically significant differences ($p<0.05$) in learning outcomes (see Part 1, above).

	Set 1	Set 2	Set 3	Set 4	Set 5
6					
6+5	391	383	440	522	644
6+5+4	603	660			1085

Table 5.5: Projected totals of annual word gains from reading sets of texts 1-5.

Figure 5.37 derives from Table 5.5 and displays the projected yearly differences in *known* word sums from children (as a group) having read one set of experimental texts as opposed to another. The intersection of Set 5 (horizontal axis) and Set 1 (vertical axis), for instance, reveals the estimated total words a child gained from reading Set 5 minus those from reading Set 1,[101] and the intersection of the Set 4 column and the Set 2 row, the additional words from the Set 4 reading experience. In the event that word gains proved significantly different under more than one definition of *known* (as identified in Part 1), then estimates appear for each. From having read exclusively Set 5, as opposed to Set 2, type texts, for example, the predicted difference in *known* word sums amounts to 261 or 425 depending upon whether the term *known* denotes word occupancy in VSAT States 6+5 or States 6+5+4. The estimated gains under either *known* notion appear somewhat variable, as expected given

[101] Recall that in no cases did the set offering the relatively massed learning result in statistically significantly more word gains.

the alternative notions of *knowing* and findings of average gains reported in Part 1. Figures range from a minimum 139 additional words per academic year from reading Set 4 as opposed to Set 2, to a maximum of 482 extra from reading Set 5 as opposed to Set 1.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					(6+5): **253** (6+5+4): **482**
Set 2				(6+5), **139**	(6+5): **261** (6+5+4): **425**
Set 3					(6+5): **204**
Set 4		(6+5), **139**			
Set 5	(6+5): **253** (6+5+4): **482**	(6+5): **261** (6+5+4): **425**	(6+5): **204**		

Figure 5.37: Projected differences in total words learned, per year, from reading texts designed to the specifications of Sets 1–5.

With the SM available (p.230), the task of deriving the annual difference in *known* word sums from reading alternative sets of texts and expressing this in standard measure units becomes an exercise of division and subtraction –dividing the SM (693) by the presumed annual sum of additional words (see Figure 5.37) from reading the set of a pair associated with most learning. The results of this computation appear in Figure 5.38. Empty cells denote pairs from which children did not gain statistically significantly more words from reading one set or the other.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					1.(6+5): **0.36 SM** 2. (6+5+4): **0.69 SM** Average of 1 & 2: **0.52 SM**
Set 2				(6+5): **0.20 SM**	1.(6+5): **0.37 SM** 2. (6+5+4): **0.61 SM** Average of 1 & 2: **0.49 SM**
Set 3					(6+5): **0.29 SM**
Set 4		(6+5)**: 0.20 SM**			
Set 5	1.(6+5): **0.36 SM** 2.(6+5+4): **0.69 SM** Average of 1 & 2: **0.52 SM**	1.(6+5): **0.37 SM** 2.(6+5+4): **0.61 SM** Average of 1 & 2: **0.49 SM**	(6+5): **0.29 SM**		

Figure 5.38: The difference in the number of target words gained as a percentage of the presumed yearly total from RR experiences, for each definition of 'known.'

Step 4; Results.

The final step in an SM analysis involves attaching pedagogical meaning to the projected learning outcomes by expressing in SM terms the additional learning from having read the set of texts from which the participant group gained significantly ($p<0.05$) the more target words.

The minimum gain (Figure 5.37) amounted to the 139 extra words annually from reading Set 4 as opposed to Set 2 type texts, the additions in this case representing 20% of the presumed yearly total (i.e. 0.20 SM) from unadapted texts. The maximum gain stands at a rather more impressive 69% (i.e. 0.69 SMs) from having read Set 5 as opposed to Set 1 type texts, and from defining *known* words as any occupying VSAT States 6+5+4. Depending upon the definition of *knowing,* word gains range from 0.20-0.37 standard measures (under the 6+5 test), to between 0.61 and 0.69 SM if *known* denotes occupancy of States 6+5+4. But what of the arguably more meaningful average gain (p.229) derived from the sum of words *known* to the State 6+5 and State 6+5+4 standard? For the pair of Sets 1 and 5, from which the participant group gained significantly more words from the latter under both notions of *knowing*, the additional learning from Set 5 type texts comes to 0.49 SM. Sets 5 and 2 represent a second pairing from which children gained significantly more words under each definition of *known,* the average gain from Set 5 type texts now amounting to an impressive 0.52 SM. The gain disparity associated with alternative notions of *known* could, it seems, often prove substantial. From Set 5 texts, for example, children benefited from a 0.69 SM gain over the sum of learned words from Set 1 scripts, assuming the 6+5+4 notion of *known*, and yet a relatively modest 0.36 SM if *known* words amount to those in States 6+5. Even the least impressive gain, however, the 0.20 standard measure increase recorded from children having read Set 4 rather than Set 2 type texts may nevertheless represent a useful addition to the average child’s lexicon. Students attaining 20% more marks in a vocabulary test, or mastering 20% more content from a lesson might reasonably expect to achieve higher class rankings or more impressive grades than peers. In the long term, the cumulative effect of such gains could prove substantial.

The pedagogical implications from reading sets of texts designed to the specifications of one set rather than another, express themselves also in time differences required for children to gain the same sum of new words. From texts designed to Set 2 specifications, for example, a child would learn an average of 2.59 words per RR session (to the 6+5 standard of *knowing*), corresponding to 383 words per year (see Table 5.5). To gain the estimated additional 261 *known* words were s/he to have read Set 5 type texts would require approximately 100 extra RR sessions (i.e. 261/2.59) with Set 2 reading materials. This amounts to 67% of the total RR hours children engage in during a single academic year. Assuming schools indeed, and as suggested, set aside 25-minutes per day for RR sessions this equates to an additional 41.6 hours of reading time. Given the 0.69 SM gain (From reading Sets 5 and 1), and with *known* words defined as those in States 6+5+4, gains appear more impressive still. For a child to learn from Set 1 type texts the sum of words s/he would from instead reading texts designed to Set 5 specifications (i.e. to make up the 482 word shortfall) would call for a further 118 RR sessions, or approximately 79% of the 148 yearly total. This amounts to 49.1 additional reading time hours otherwise available for non-RR purposes. A full listing of predicted additional RR sessions required for children to gain a comparable number of words from the set of texts less optimal for gain of each possible pairing appears in Figure 5.39 below.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					(6+5): **97.3** (40.5 hrs.) (6+5+4): **118** (49.1 hrs.) Average of 1 & 2: **107.6** (44.8 hrs.)
Set 2				(6+5): **53.7** (22 hrs.)	(6+5): **100** (41.6 hrs.) (6+5+4): **95.5** (39.7 hrs.) Average of 1 & 2:**97.75** (40.65 hrs.)
Set 3					(6+5): **101** (42.1 hrs.)
Set 4		(6+5**):** **53.7**(22 hrs.)			
Set 5	(6+5): **97.3** (40.5 hrs.) (6+5+4): **118** (49.1 hrs.) Average of 1 & 2: **107.6** (44.8 hrs.)	(6+5): **100** (41.6 hrs.) (6+5+4): **95.5** (39.7 hrs.) Average of 1 & 2:**97.75** (40.65 hrs.)	(6+5): **101** (42.1 hrs.)		

Figure 5.39: Additional number of daily RR sessions (**bold**), and hours (hrs.), required to gain the yearly sum of known words from the set of a pair offering the more distributed learning opportunities.

The estimates in Figure 5.39 supply no indication of quality of gains, i.e. depth of understanding, however. From a teacher's standpoint, quality depends upon the distribution of gained words among VSAT defining the *known* word class. Since the current study did not employ the VSAT to estimate *known* word gains from reading unadapted texts, what children knew of words Nagy et al. (1987) might have identified as *known* remains unclear, the authors providing only broad details of what *knowing* implies (See p.230). Without evidence of how the Nagy et al. (1987) *known* words might have dispersed among VSAT states, SM figures potentially allow for very different interpretations. A child who understands 10 words to the standard of VSAT State 6, for example, possesses more knowledge of those words than another who *knows* them to the standard of VSAT State 4, just as *knowing* a word to the standard of State 5 implies a higher knowledge than were s/he to *know* the same word to the standard of State 3. Interpreting SMs, then, necessarily remains subjective, the more so that the Nagy et al. (1987) test of *knowing* departs from the VSAT defined alternatives of (a.) word occupancy of States 6+5 or, (b.) occupancy of States 6+5+4. Should Nagy et al. (1987) have recognized a more lenient (i.e. easy to satisfy) test of *knowing* than either VSAT alternative, then the additional words from a set of a two-set combination will amount to less (numerically) in SM terms, since the VSAT yields a lower *known* word total than would Nagy et al.'s (1987) test procedure. The depth of those gains, however, exceeds that from unadapted texts given the lesser understanding that suffices for Nagy et al. (1987) to assign words a *known* designation. The 'opposite' also applies. Should the Nagy et al. (1985) test of *known* prove more stringent than the VSAT alternatives, then the total word gains from a set now become more impressive in SM terms; depth of understanding, however, will amount to less than from unadapted texts since the VSAT will identify some words as *known* that the Nagy et al. (1987) test would have rejected as such. This becomes important when interpreting SM data: Word gains expressed in SM terms could prove familiar to a deeper or shallower degree relative to those from authentic texts depending upon (1.) how comparable are the tests of *known* word for computing the sum from unadapted texts, and (2.) the actual differences in *known* sums from reading the sets of a pair of interest. At the present time, the ambit of what 'knowing a word' entails remain short of ‘settled.’

<u>Conclusion 55</u>: Under a definition of *known* words that includes within the *known* category those words assigned to either VSAT States 6+5 or 6+5+4, the differences in the degree of spaced learning a set of texts afforded accounted for, potentially, word gains of between 20% and 69% of the sum readers might have gained from RR sessions with unadapted texts (per year).

<u>Conclusion 56</u>: In those cases, where the sum of *known* words from reading a pair of sets revealed significant differences under more than one definition of *known*, the choice of what amounts to a *known* word substantially influences how impressive the gains deriving from the spaced learning condition.

To assume that SM findings shed clarifying light on spaced learning rests upon the study’s use of texts that differ only in regard to the target word presentation time they afforded (See Chapter 4). That children indeed gained significantly more words, and to various standards of understanding, therefore legitimately implies a spaced learning impact. How impressive a teacher might regard such gains in pedagogical terms depends upon the receptive and/or productive skills s/he happens to rate most highly. For a class of young EAL students requiring large receptive vocabularies for social interaction and curriculum access, shallow familiarity with a large sum of words to the VSAT States 5+4 standard may seem somewhat more desirable than a deep knowledge (e.g. *knowing* words to the VSAT State 6 standard) of rather fewer. The opposite might apply, however, if preparing students to participate in ‘speech giving’ or report writing where vocabulary choices and formal accuracy become more critical. That children might fail to gain significantly more words of the type they could supply in semantically and syntactically well-formed sentences (words familiar to the State 6 standard) from a hypothetical year’s worth of spaced learning does not imply no learning of words to this standard, however; some such learning would indeed occur, as confirmed by findings reported in Part 1 that reveal State 6 word gains even from children having read Set 1. The current study establishes, rather, that manipulating the time intervals between encountering the same novel word fails to translate into statistically significant ($p<0.05$) differences in word gains assuming the strict VSAT State 6 notion of what ‘*known* word’ implies. For teachers aiming to develop students’ high level productive word knowledge, direct vocabulary instruction would likely prove rather more fruitful than RR per unit of instructional time.

The one lexical competence in which children most likely demonstrated an improvement from texts optimized for spaced learning stands as the ability to supply good synonyms of formerly unknown words –the competence VSAT State 5 seeks to capture (Section 5.8.3). The evidence for this seems compelling, as we have seen (Section 5.12.3). In contrast, the probability that more, or less, spaced learning affects the sum of words children gain to the receptive State 4 standard comes across as conspicuously low. Both pairwise McNemar tests and the vote-count/sign test, indicated a stubborn resistance of State 4 sums to presentation time differences.

5.20 Significance of word gains differentiated by class (Research Question 2)

This section moves beyond gross word gains to analyze differences in *known* word sums differentiated by lexical class, the issue of concern to Research Question 2. As in the gross sum analysis (above), the section applies the standard measure construct to express gains in terms of pedagogical significance. The methodology forms the subject of Section 4.22.1.

Step 1. *Apportioning the presumed yearly gross sum of learned words from unadapted reading materials.*

The Nagy et al. (1987) probability estimate of novel word learning represents a general 'ballpark' figure –a figure that sufficed given that particular study's objectives but which does not acknowledge some word classes as relatively more learnable than others. To derive SMs for a single lexical class requires extracting from the gross learned word total, the gains of each lexical category of interest. Only this yields the relevant yearly *known* content word sums (i.e. SMs), and allows for deriving probability estimates of learning a word of that class from a single encounter. As a workable, and reasonable, approximation, the study assumes that the percentages of nouns, verbs, adjectives and adverbs (as reported in the *British National Corpus*) among the first 4,000 most common English language lemmas corresponds to the percentages of words of these classes children gain from RR sessions. That this is a sensible supposition derives from two observations. First, that children (and especially EAL students) gain much vocabulary from texts as opposed to conversational exchanges, with reading contributing substantially to vocabulary development beyond the first few thousand most common words (Nation, 2000). And, second, the 4,000 lemmas likely represent a significant 87.6% of the words in scripts generally (Carroll, Davies & Richman, 1971), and a somewhat higher percentage still of words children and adults employ during informal social interaction.

To identify the BNC proportions of nouns, verbs, adjectives and adverbs involved summing the lemmas for each class in the BNC corpus within the first 4000 frequency-ordered entries. This tallying revealed nouns comprised roughly 50.4% of word types (2,019/4,000), verbs 20.4% (817/4,000), adjectives 16.5% (660/4,000) and adverbs just 7.4% (297/4,000). The remaining 5.3% of lemmas included prepositions, articles, modals, conjunctions and interjections, among several others. With these proportions at hand, computing the standard measure can now proceed as in the same manner employed for estimating the yearly total (undifferentiated by lexical class) *known* word sums (see Section 5.4.2, above). For the class of interest, the following formula yields the SM total:

Standard measure of word type 'x' = (total words gained i.e. t.w.g) (*proportion of words of type 'x'*)

The term t.w.g refers to the total words learnt to a criterion standard from one year of RR with unadapted texts, while the *proportion of words of type 'x'* denotes the percentage of words of a class of interest of the total first 4,000 most common lemmas in the BNC corpus. Assuming, for example, children gained 150 words per year overall, then applying the formula gives a breakdown of 24.75 adjectives (0.165 x 150), 75.60 nouns (0.504 x 150) and 11.1 adverbs (0.074 x 150). The sums of each word type (i.e. 24.75, 75.60, and 11.1) in this hypothetical instance provide the relevant SM figures for input to Step 2 (below).

Employing BNC derived proportions as proxies for children's word gains (by word class), and to derive learning probabilities rests on several assumptions. Clearly, the actual percentages extracted from the 4,000 lemmas may vary somewhat according to the writer's intended audience, his or her writing style, and by genre. Somewhat different figures, for example, could conceivably have arisen from a corpus comprised of expositories, playscripts or newspaper reports rather than narratives of the sort the current study employs. Moreover, not all words in the BNC corpus derive from written materials, the compilers having referenced transcripts of spoken English such as spontaneous conversations, as well as recordings of meetings and events. More generally, deriving SMs from proportions unavoidably becomes less reliable to the extent narratives comparable to those of the present study's experimental texts (designed for recreational reading among nine-year-olds) do not appear in the BNC data source –realistically such texts comprised but a small proportion. On the other hand, such methodological concerns may prove largely misplaced. While oral language indeed figured in the BNC data pool, written scripts contributed a substantial 90% of the 100,000,000 word corpus. The scripts themselves derive from an eclectic range of materials including newspapers, technical journals, fiction, and narratives –in other words, a representative sample of adult and child literature. Despite an obvious bias towards adult texts, the BNC corpus reflects the variety of classroom reading matter from which children might select during RR sessions, with narratives comprising a substantial proportion of the BNC's lemma total. Indeed, that the BNC frequency lists derive from different text types should, arguably, only frustrate SM derivation if the relative sums of content words derived from these types differ from the sums found in young children's reading materials. Most works dealing with vocabulary development implicitly deny any such presumption. Neither Nation (2001), or Nation and Waring (1997), for example, differentiate by genre in their estimations of the total number of words readers need *know* for adequate comprehension, stressing instead the general value of common occurring vocabulary.

<u>Deriving the SMs</u>

Applying the SM formula (above) and given the previously estimated 693 novel word gain per year, then children's yearly word gains amount to 349 nouns (50.4% of 693), 141 verbs (20.4% of 693), 114 adjectives (16.5% of 693) and 51 adverbs (7.4% of 693). These figures, therefore, will serve as annual SMs (from regular RR texts) against which to compare word gains from reading alternative pairings of sets employed in the current study. The remaining words from the 693 yearly sum belong to non-content word classes (prepositions, conjunctions, determiners etc.) and account for

5.3% of the total.

Step 2. Expressing gains in SM form.

To express word gains in SM form calls, first, for deriving children's presumed yearly gains of nouns, verbs, adjectives and adverbs from texts designed to the specifications of Sets 1 to 5. This involves two steps: (1.) identifying the annual sum of novel words (undifferentiated by lexical class) learned from reading texts constructed according to the specification of that set of interest, and (2.) establishing the proportions of that total consisting of the relevant content word type (i.e. noun, verb, adjective, adverb). The annual sum (step 1) figure comes from of gross *known* word sums identified in Section 5.4.2 which reported totals of between 391 and 644 given the VSAT State 6+5 notion of *knowing*, and 603 and 1,085 if limiting *known* words to those in States 6+5+4. The second critical figure –i.e. the proportion of these sums of a lexical class of interest– derives from multiplying the gross *known* word sum from reading a set of texts by the presumed percentage of that sum children gained of the relevant lexical class; the percentages, as noted, correspond to the proportion of words of a class among the first 4,000 most common words in the BNC corpus.

Total words of type 'x' from reading set 'y'= (Sum of *known* words from set 'y,' for a definition of *known*) x (BNC proportion[102] of words of type 'x')

Applying the formula gives the hypothetical number of *known* words by word type over a single academic year, the relevant figures appearing in Table 5.6. From texts designed to the specifications of Set 2, for example, and with *known* words limited to those occupying VSAT States 6 or 5, the projection reveals a gain of 193 nouns (i.e. 0.504 x 383).[103] For Set 5, with *known* words as those in States 6 or 5, the prediction is of a 325 noun gain. For verbs, the projected total equals 123 words from Set 1 type texts if *known* denotes word occupancy of States 6+5+4 (0.204 x 603), and a rather more substantial 221 from reading Set 5 (0.20 x 1,085). Blank cells in Table 5.6 indicate

[102] See p. 219 for the derivation of these figures.
[103] The figures for the gross sums of known words from reading texts designed to the specifications of one or other set employed in the current study appear in table 5.5, above.

combinations of sets of texts, and definition of *known* word, for which tests (See Part 1) failed to reveal statistically significant ($p<0.05$) differences in learning words of the same class from having read a set of a pair.

Figure 5.40 derives from Table 5.6 and displays the estimated yearly difference in *known* word sums by lexical class for each pairing of sets (see Part 2) from which children gained significantly more words from reading one or the other. From reading Set 5 type texts a child would, for example, gain 86 more verbs than from texts designed to the specifications of Set 2 (*known* words as those in States 6+5+4), and 104 extra nouns over those from reading Set 3, with *known* words restricted to those in VSAT States 6+5., Figure 5.41 draws upon findings in figure 5.40 along with the standard measure sums to express differences in *known* word sums in SM form for both nouns and verbs, the two content word classes of the four associated with significantly more learning under spaced conditions. The following formula provided the proportion of the SM figures that appear in each cell:

$$\left(\frac{\text{difference in the annual sum of ‘known’ words of type ‘x’ from reading set ‘x’ and set ‘y’}}{\text{total of ‘known’ words of type x gained during one year from unadapted texts}} \right) \text{x } 100$$

	Set 1	Set 2	Set 3	Set 4	Set 5
6					
6+5		Nouns: 193 (av.1.30)	Nouns: 221 (av.1.49)		Nouns: 325 (av.2.19) Verbs: 131 (av.0.88)
6+5+4	Verbs: 123 (av.0.83)	Nouns: 333 (av.2.25) Verbs: 135 (av.0.91)			Nouns: 546 (av.3.69) Verbs: 221 (av.1.49)
Av. of 6+5 & 6+5+4		Nouns: 234 (av.1.77)			

Table 5.6: Sums of known word by lexical class from reading alternative sets of texts (av. = average number of words learned per RR session).

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					6+5+4, v, **98**
Set 2					1. 6+5, n, **132** 2. 6+5+4, n, **213** Average of 1 & 2: n, **172.5** 6+5+4, v, **86**
Set 3					6+5, n, **104**
Set 4					
Set 5	6+5+4,v, **98**	1. 6+5, n, **132** 2. 6+5+4, n, **213** Average of 1 & 2: n, **172.5** 6+5+4, v, **86**	6+5, n, **104**		

Figure 5.40: The difference in totals of 'known' words from reading pairs of sets.

	Set 1	Set 2	Set 3	Set 4	Set 5
Set 1					(6+5+4), v, **0.70 SM**
Set 2					(6+5), n, **0.38 SM** (6+5+4), n, **0.61 SM** Average of 1 & 2: n, **0.49 SM** (6+5+4), v, **0.61 SM**
Set 3					(6+5), n, **0.30 SM**
Set 4					
Set 5	(6+5+4), v, **0.70 SM**	(6+5), n, **0.38 SM** (6+5+4), n, **0.61 SM** Average of 1 & 2: n,**0.49 SM.** (6+5+4), v, **0.61 SM**	(6+5), n, **0.30 SM**		

Figure 5.41: Differences in totals of learned words expressed in terms of standard measures.

Results:

The most striking indication from Figure 5.41 is the magnitude of the SM scores. The 0.70 SM gain in verbs from reading Set 5 as opposed to Set 1 type texts (with *known* words as those in VSAT States 6,5, or 4) amounts to just marginally short of three-quarters the sum of this word type children might gain from unadapted texts per year. The least substantial difference in *known* sums, the 0.30 SM additional nouns from Set 5 rather than Set 3 type texts (with *known*, as word occupancy of States 6+5), represents almost one third the presumed words of this class gained annually from unadapted scripts during RR opportunities. The average noun gain –the average of the figures for 6+5 and 6+5+4 tests of *knowing,* and an arguably more meaningful measure than either alone, comes to 0.49 SM corresponding to almost half the predicted yearly sum of nouns learned from regular RR texts. The savings in instructional time from optimally designed texts would, then, appear substantial. With *known* words as those occupying States 6+5+4, for a child to 'make up' the 213 noun shortfall from reading Set 2 as opposed to Set 5 type texts calls for a predicted additional 95 daily RR sessions (per year) with Set 2 type materials; that is, 64% more RR sessions than the 148 presumed (academic) yearly total.[104] The 0.30 standard measure gain in nouns from reading Set 5 in preference to Set 3 type texts corresponds to a difference in the *known* sum of 104, implying 69.7 extra RR sessions for children to gain as many nouns from Set 3 type texts as from those designed to Set 5 specifications. A full listing of the additional RR sessions and instructional time (in hours) required to gain comparable totals of *known* words of a class from the set of a pair presenting the more massed target word encounters appears in Figure 5.42, below.

Conclusion 57: A statistically significant ($p<0.05$) difference in the sums of *known* words of a particular class from the more distributed learning a set of texts afforded could amount to gains of between 98 (minimum) and 213 (maximum) additional words per year, per child.

Conclusion 58: For nouns, the maximum difference in *known* totals (in the event that the totals were significantly different) from reading sets of texts designed to the specifications of those employed in the present study amounted

[104] The average sum of words gained per RR session from unadapted texts, assuming 148 sessions per year are: nouns (av. 2.3), verbs (av. 0.95), adjectives (av. 0.77), and adverbs (av. 0.34).

to 132 under the 6+5 test of *knowing* (from reading Sets 5 and 2) and 213 from reading Sets 5 and 2 given the States 6+5+4 definition of *known* word.

<u>Conclusion 59</u>: For verbs, the maximum difference in *known* word totals (should totals have proven significantly different as established in Part 1) from reading sets of texts designed to the specifications of those the present study employed ranged from 0.61 to 0.7 standard measures (only under the 6+5+4 definition of *knowing* did sums differ significantly).

	Set 1	Set 2	Set 3	Set 5
Set 1				(6+5+4), v **119**, (49.1 hrs.)
Set 2				1. (6+5), n, **101**, (42.08 hrs.) 2. (6+5+4), n, **95** (39.58 hrs.) Average of 1&2: **97**, (40.41 hrs.) (6+5+4), v, **94.5** (39.37 hrs.)
Set 3				(6+5), n, **69.7** (29.04 hrs.)
Set 5	(6+5+4), v, **119,** 49.1 hrs.)	1.(6+5), n, **101,** (42.08 hrs.) 2. (6+5+4), n, **95,** (39.58 hrs.) Average of 1&2: **97**, (40.41 hrs.) (6+5+4), v, **94.5** (39.37 hrs.)	(6+5), n, **69.7,** (29.04 hrs.)	

Figure 5.42: The number of extra RR sessions (bold) required to gain the same number of words from texts least optimal for vocabulary gain as those from the set most optimal.

5.21 A summary of gains by word class

The study finds children did not gain statistically significantly more words of any lexical class to the VSAT State 6 criterion of *knowing,* and only rarely more words of any lexical category under either of the two less demanding tests the study employs (i.e. word occupancy of States 6+5, or 6+5+4). Should the participant group have learned more words of a class from a set of a pair then that class was exclusively either noun or verb.

While the SM values present gains in more meaningful terms than bare totals they still leave ample scope for disagreement. A teacher might express satisfaction with the time saving and fewer RR sessions to gain the same sum of additional words from texts optimally designed with spaced learning in mind and yet harbor serious concerns regarding the improbability of statistically significant differences in learned word sums across all content word types (see Section 5.18). Adjective and adverb insensitivity to presentation time variation might seem particularly unfortunate given the information bearing role these word types play. Even among nouns and verbs, how impressively spaced learning contributes to vocabulary growth depends upon one's understanding of what amounts to a *known* word from among several reasonable possibilities. For nouns, for example, assuming children read Set 5 types texts as opposed to Set 2, whether one denotes *known* as occupancy of VSAT States 6+5 or, more loosely, as occupancy of States 6+5+4 implies gains of 38% or 61% respectively of the annual noun total from unadapted texts i.e. 0.38 and 0.61 SMs (see Fig. 5.41, p.243). The average of the States 6+5 and States 6+5+4 SMs, the rather more meaningful measure of learning, stands at a credible 0.49 SMs. For those teachers striving to develop learners' productive vocabulary knowledge, on the other hand, the spacing effect might seem to offer little, and especially so given the technical challenges, sheer effort, and research that textual adaptation entails. It is not obvious that the time and planning involved in modifying reading materials yields a reasonable return in children's *known* word gains.

5.22 A speculative note on gains compared to those from reading unadapted texts

The current study compares word gains from reading one experimental set of texts with an alternative set. It has not as yet explored, or commented upon, differences in learning from reading experimental sets on the one hand, and unadapted on the other. Whether children learn more from adapted than authentic RR materials receives attention here given the insights this sheds on whether manipulating time intervals between novel word reencounters could prove potentially worthwhile. While the comparison reveals nothing of associations between distributed encounters and novel word gains (see the discussion of confounding variables, Chapter 4) this is of little concern to teachers interested in vocabulary uptake as such, rather than those more intrigued by constructing models of the learning process.

The sum of words participants in the present study gained per year from reading Set 5 type texts –i.e. texts from which children made the most impressive learning– appears in Table 5.7 below (Column 1) for each definition of *known* associated with significant ($p<0.05$) differences in target word uptake. The figures rest upon the assumptions informing the discussion in Section 5.4.2, namely: (1.) that a child encounters approximately 8,158 novel words per year, (2.) that s/he engages in daily RR sessions of around 25 minutes duration, (3.) that average reading speed amounts to approximately 147 wpm, and (4.) that Nagy et al.'s (1987) probability of word gain per encounter (i.e. 0.085) reasonably applies to the 9-year-old Thai L1 children participating in the current project. The annual word gain from unadapted texts appears in Column 2 (693 words), the figures being the product of the 8,158 unknown words encountered per year and Nagy et al.'s (1987) probability of novel word learning (see Section 5.4.2, above).[105] The Nagy et al.(probability) recommends itself for the reasons reported earlier in the discussion of gross *known* word totals (Section 5.19.1) –that if derives from observations of grade 3 school children's experiences with narratives, and presumes reader familiarity with the concept the novel word denotes.

The third column of Table 5.7 displays the estimated sum of additional words assuming participants in the present study to have read Set 5 type texts as opposed to unadapted RR scripts over an academic year –i.e. the sum of Column 1 minus Column 2. The figures for average gains per RR session amount to the product of the hypothesized annual sum of novel words learned (the aforementioned 8,158), and the probabilities of word gain from a single encounter (see Table 5.4, p. 231). The fifth column of Table 5.7 displays the estimated sum of words gained from unadapted texts as a proportion of those from authentic reading materials assuming the same one year time period.

The difference in presumed gains from having read authentic as opposed to adapted texts

[105] This 693 sum is the same figure that served as the standard measure against which to quantify the magnitude of spaced learning gains from reading alternative sets of texts (Section 5.4.2).

clearly varies quite substantially, depending upon the sense of *knowing* one acknowledges. While children would *know* 49 words less from Set 5 type texts than from unadapted texts if *known* denotes word occupancy of VSAT States 6+5, a child's gain comes to a relatively impressive 392 words under the less restrictive test that recognizes *known* words as any assigned to States 6+5+4. Expressed as percentages, and projecting over one academic year, children would have learned between 93% (the 6+5 test) and 156% (the 6+5+4 test) of the novel words appearing in authentic RR texts had they instead read adapted scripts supplying the same novel vocabulary terms. Table 5.8 reports the average of word gains under the State 6+5 and State 6+5+4 notions of *knowing*, conflating the two into a composite notion that arguably corresponds more closely to the Nagy et al. (1987) understanding of *known* from which the authors derived their 0.085 learning probability estimate (see Section 5.4.2).[106] This average figure yields a prediction of 171.5 extra words per year from Set 5 type reading materials. In percentage terms the 'average' Year 4 child gains 124.7 % of the presumed words learnt had s/he read purely authentic reading materials.

	(1) Word gains from Set 5-type texts.	(2) Word gains from unadapted texts.	Difference between (1) and (2).	Set 5 'Known' words as a percent of unadapted text gains.
6+5	644 (av: 4.35)	693(av: 4.68)	-49	93%
6+5+4	1,085 (av: 7.33)	693(av: 4.68)	392	156.5%

Table 5.7: Differences in word gains from reading adapted (Set 5) and unadapted texts (av=average number of words gained per RR session).

106 That is, the tests from which Nagy et al. (1987) computed the probability of word gain from a single encounter.

The savings in both the number of RR sessions and instructional time (hours) from textual adaptation appear in Table 5.9 for each definition of word *knowing*, along with averages of the two (Details of the computation appear in Section 5.4.2). Assuming 148 RR sessions per year, and 25 minutes per session, children will require an additional 11.26 RR sessions beyond the 148 total to gain the same sum of words (to the 6+5 standard of *knowing*) from Set 5 type texts as they would from unadapted over the course of one school year. The instructional time saving amounts to 4.69 additional hours of RR (25 x 11.26), or a 7% increase in the number of annual RR sessions. Under the 6+5+4 notion of *knowing* the benefits from Set 5 type appear more substantial still, corresponding to a time saving of 36.6 hours or 87.86 RR sessions. The average of the gains given the 6+5 and 6+5+4 tests of *known* implies children would require approximately 36.7 additional RR sessions per year with authentic RR scripts to gain the same sum of words learned from texts designed to Set 5 specifications. Assuming a child engages in 4 daily RR sessions per week, this amounts to an extra 9 weeks' reading unadapted RR materials. The findings go some way towards recommending adapted texts as part of a comprehensive strategy to resolve vocabulary deficits –all the more so should children experince little exposure to English outside of classroom settings or only limited opportunities to interact with English L1 users generally.

(1) Word gains from Set 5 type texts (average of sums learned under the State 6+5 and States 6+5+4 tests of *known* word.	(2) Word gains from unadapted texts.	(3) Difference between 1 and 2.	(4) Set 5 'known' words as a percent of unadapted text 'known' word gains.
864.5 (av. word gain per RR session: 5.83 words).	693 (av. word gain per RR session: 4.68 words).	171.5 words.	124.7%

Table 5.8: Additional number of words gained from reading adapted texts.

	Additional RR sessions.	Additional time.
6+5	11.26 sessions with adapted texts.	4.69 hrs. with adapted texts.
6+5+4	7.86 sessions with unadpated texts.	36.56 hrs. with unadpated texts.
Average	36.7 sessions with unadapted texts.	15.3 hrs. with adapted texts.

Table 5.9: Additional number of RR sessions, and instructional time to gain the same number of words from reading the text associated with the lesser learning.

Table 5.10 presents projections of differences in word gains for nouns and verbs, the two classes of the four for which McNemar tests (Part 1) indicated children learned statistically more novel words under spaced learning conditions. The figures in Column 2 refer to Set 5 type texts and derive from the yearly noun and verb word gain totals that appear in Table 5.6 (above) –the same sums that served (Section 5.20) when comparing gains of these classes over one academic year. Column 3 depicts projections for noun and verb learning from unadapted texts based upon the annual *known* word total given (1.) Nagy et al.'s (1987) 0.085 estimated probability of word learning from a single encounter, and (2.) the proportion of total gains of each class implied by the percentage of nouns and verbs in the 4,000 most commonly occurring BNC lemmas (see Section 5.20). Column 4 displays the difference in word gains (i.e. Column 2 – Column 3) by word type from children having read authentic or adapted texts during one academic year of RR opportunities. Figures in the final column, Column 5, indicate the sum of words gained from reading Set 5 as a proportion of those from unadapted texts (i.e. gains in SM units).

What do the data reveal? Under the least demanding test of *known* (occupancy VSAT States 6, 5 or 4) the projection for nouns amounts to 197 additional words from Set 5 type texts, this corresponding to a 56.4% increase over the total nouns from unadapted reading material total. For verbs, the gain proves marginally lower at 56.7%, translating to an 80 word addition. The apparent advantage we see in modified texts, however, no longer persists if we restrict *known* words to those in States 6+5. Under this, somewhat stricter, definition of *known* word the prediction now becomes a 25 noun deficit (a reduction of 7.1%) from Set 5 type scripts (324 nouns as opposed to 349), and 10 less verbs. As with the data for gross sums (the total of nouns, verbs, adjectives and adverbs), the average of words gained under the 6+5 and 6+5+4 notions of *known* once again indicates an adapted text advantage. For nouns, gains now amount to 86 words, or 24.6% the yearly sum presumed from unadapted RR materials, and for verbs an additional 35, or 24.8% of that sum –i.e. 0.248 SM. Based on the average number of estimated words of each class learned per RR session (see Table 5.10), then remedying the noun deficit, should children exclusively read authentic texts, would require approximately 36.5 further RR sessions per academic year (an increase of 24.6%).[107] Expressed as instructional time, the time saving from employing adapted texts amounts to approximately 15.20 hours, assuming 25 minute RR sessions. The additional number of sessions per year for children to gain the same number of verbs from authentic, as opposed to Set 5 type texts, amounts to roughly 36.8, i.e. 15.3 extra hours.

Definition of known word.	(2) Word gains from Set 5-type texts.	(3) Word gains from unadapted texts	(4) Difference between 1 and 2.	(5) Set 5-type word gains as a proportion of unadapted text gains.
State 6+5				
Nouns	324	349	-25	92.8%
verbs	131	141	-10	92.9%
State 6+5+4				
nouns	546	349	197	156.4%
verbs	221	141	80	156.7%
Average of the 6+5 and 6+5+4 gains.				
Nouns	435 (av. 2.93)	349 (av. 2.35)	86	124.6%
Verbs	176 (av. 1.19)	141 (av. 0.95)	35	124.8%

Table 5.10: Differences in words gained from reading adapted (Set 5) and unadapted texts (av. =average number of words gained of the relevant class per RR session).

5.23 Conclusions and Limitations, Part 3

SM findings shed no light on aspects of literacy competence other than the depth and breadth

[107] Based on the average number of words gained per RR session reported in Columns 1 and 2.

of word meaning. They tell us nothing, for example, of whether more distributed learning impacts upon a child's writing style, his or her grammatical accuracy, spelling skills, or punctuation. Nor do the figures hint at the magnitude of word gains had children received direct vocabulary instruction –a well-attested albeit apparently uncommon source of vocabulary development in K-12 settings. To these limitations should be added several methodological concerns. While SMs helpfully place word gains into a pedagogical context of sorts, conclusions on learning outcomes must remain necessarily speculative given the limited data available. As noted, projected word gains and SM derivations rest upon a raft of assumptions, some rather more tenable than others. The Nagy et al. (1987) probability (0.085), for example, derives from English L1 learners attending American schooling and may not similarly apply to privately educated Thai EAL pupils for whom vocabulary gains from RR have not as yet attracted much research interest. The more these two probabilities –the *supposed* (i.e. Nagy et al. 1987) and *actual*– diverge, the less robust the SM construct for expressing word gains as a proportion of those from regular RR sessions. Nor can one ignore the critical topic of memory decay. Children will inevitably fail to retain a proportion of word meanings learned during RR sessions without regular reinforcement. Inevitably, gains derived from VSAT tests conducted promptly after a learning opportunity will overestimate what a child goes on to retain in long term memory. Factoring in 'forgetting' is nothing short of essential in the search after reliable estimates of learning gains from RR sessions. A child having gained knowledge of a word from reading experiences early in the academic year will unlikely retain that knowledge without having met the same word on multiple occasions thereafter. Not least, and as noted in Part 1 (150-166), any findings applicable to a participant group may have little relevance to individual members.

5.24 Individual learning outcomes

That participants collectively gained words from their reading sessions does not deny scope for substantial variation in learning among individual children. The data reported in Sections 5.4.3, 5.5.3, and 5.6.3 indicates this holds true of all definitions of *known*, and irrespective of which sets of texts supply the raw data for comparisons. Participant 1, for example, learned all four words from reading Set 4 (*Known* = occupancy of States 6+5+4), while participant 11 gained only one; participant 3 gained four words from Set 3 (with *known* as words in States 6+5) and participant 7, zero. Appendix 2 reveals numerous other examples of just such score disparities. The same variability also presents when we consider words differentiated by lexical class. Fourteen participants gained a noun to the State 6 VSAT standard from reading Set 1, while 6 did not *know* the relevant target noun even with the class of *known* words expanded to include occupants of States 6+5+4. From reading Set 5, five children *knew* an adjective to the State 6 standard, and yet for 7 others the word remained unfamiliar under each test of *knowing* the study employed.

The average of *known* word gains for the participant group, and the corresponding figure for individual members, often diverged, making the group figure a poor predictor of any one student's learning performance. The mean number of *known* words from reading Set 5 under the least restrictive notion of *knowing* (word occupancy of either VSAT State 6,5 or 4) amounted to 3.29 words (Section 5.6) while for any one child this could extend from a high of four words to a low of two.

Conclusion 60: For any target word, the difference in what a child *knew* of its meaning relative to that of his/her peers could be large, with some children *knowing* a word to a standard permitting productive use and others unable to provide even a broad synonym to satisfy the test for occupancy in State 4.

Just as *known* word sums could vary among children, so too, a child's relative ranking among his or her peers. General statements of the form that student 'x' or 'y' is a 'poor' or 'weak' learner clearly mean little without qualifying the sense of *knowing,* and the particular sets of texts the reading involved. Participant 24, for example, proved the least capable learner from reading Set 1 under the States 6+5+4 based definition of *known* and yet the most able from reading Set 5. Participant 11 gained zero words from Set 1 (*known* =word occupancy of State 6) yet ranked equal first based upon Set 5 scores. Indeed, correlations of children's ranks for the same definition of *known* word appear generally unimpressive. The results from pairwise Spearman rank order test of children's placements derived from words familiar to the VSAT State 6 revealed the highest correlation between Sets 2 and 3 (-0.67; p=0.000); seven of the pairwise comparisons yielded negative values, and three positive; r_s (rho) values in all cases lay within the 'weak' to 'very weak' range given Landis and Kosch's (1977) interpretive guidelines. Under the stricter 6+5 test of *knowing* the Spearman tests returned three negative correlations and seven positive with r_s (rho) again falling within the 'weak' to 'very weak' range; the highest correlation derived from comparing Set 2 and 3 rankings, an $r_s = 0.35$. For the 6+5+4 based definition of word *knowing*, negative correlations numbered six, and positive just three. The 'strongest' correlation emerged when comparing Sets 3 and 5, albeit again weak at $r_s = 0.222$.

<u>Conclusion 61</u>: The sum of words a child gained from reading any one set of texts (and his/her ranking relative to peers based on those rankings) for a particular definition of *known* failed to reliably predict (a) how many words s/he gained from reading another set of texts, or (b) his/her ranking in relative to peers in terms of word gains from that set.

More generally, in regard to performance relative to peers:

<u>Conclusion 62</u>: For any one set of texts, a child's ranking in terms of word gains, relative to his/her peers could vary by several places depending upon the particular test of word *knowing* one applies.

This variability in children's word gains suggests the optimum time interval between target word presentations uniquely differs according to the child in question. While participants generally learnt more words from Sets 5 and 4 than Sets 1 and 2, Set 5 did not invariably prove the most learning conducive; Appendix 2 reveals numerous illustrations. Participant 11, for instance, gained three words (*known* = 6+5+4) from reading Set 1, only to learn one from reading Set 4. Participant 23 gained three words from Set 1 (*known* = word occupancy of VSAT States 6+5) but two from Set 5, while participant 18 gained three words from Set 3 and none from Set 5 (*known* = occupancy of VSAT State 6). Indeed, for any definition of *known,* one or more children gained more target vocabulary from Set 1 than from Sets 4 or 5 despite the participant group gaining most words under the relatively spaced target word presentation time Set 5 provided. Why a child learned more words from one set of texts than did a peer raises questions of individual learning capacity as much as the impact of time intervals between learning occasions. The issue receives attention in the following chapter.

<u>Conclusion 63</u>: **The study provides some evidence that the optimal degree of spaced learning (for any given notion of word *knowing*) displays variability depending upon the child in question.**

<u>Conclusion 64</u>: **It is possible –or at least implied in the data– that what amounts to an optimal degree of distributed learning for any one child may depend on the test of word *knowing* employed.**

The number of children displaying a progression of more (or equal) learning from reading numerically higher ordered sets ranged from a minimum of four (*known* = States 6+5) to six (*known* = States 6+5+4). Under no definition of word *knowing*, however, did a child exhibit the reverse condition of consistently fewer (or equal) learned words from a set presenting target words under a more massed than distributed presentation (i.e. a pattern of least gain from Set 5, rather more from Set 4, more still from Set 3 and so on). Leaving aside data from Sets 2 and 4 (that is, introducing a longer difference in the time over which word presentations occurred), the number of participants gaining additional words from the higher numbered set of a pair amounts to 5, for VSAT State 6, 4 for States 6+5, and 8 if *known* words comprise those in either State 6, 5 or 4. The figures support a tentative conclusion of more robust associations between word gains with ever broader and more inclusive definitions of word *knowing*.

<u>Conclusion 65</u>: **For every combination of sets of texts and definition of *known* associated with the participant group gaining statistically significant ($p<0.05$) difference in total *known* words, individual learners could gain more words from the set associated with the lesser gains.**

<u>Conclusion 66</u>: **(Notwithstanding Conclusion 65), the more expansive the definition of *known*, the less the number of children gaining more target words from a set exposing them to the more massed learning condition than another.**

The number of 'no gainers,' i.e. children having failed to master any words from reading a set of texts received attention in Sections 5.4.4, 5.5.4 and 5.6.4 and offers an alternative perspective on spacing effects that focuses not so much on how much learning occurred but upon whether any learning occurred at all. Data on zero uptake, at first glance, may seem to reveal little of pedagogical efficacy since it says nothing of how many words a child mastered, or the depth of understanding of those gained. Nor can one discount the confounding possibility that children may self-claim no familiarity with a target word during VSAT testing and yet objectively display some knowledge of which they are unaware, nonetheless. This, emerged as one of the more surprising conclusions of Durso and Shore's (1991) study with tertiary level participants (Section 3.2). These concerns aside, data in regard to no learning does usefully establish the proportion of students for whom a teaching practice has no effect on attainment, as opposed to one that potentially has some impact. What amounts to, or signifies, 'zero knowledge,' however, lies open to alternative and reasonable interpretations. In the current study, the candidate VSAT states designative of zero understanding comprise, from least to most controversially, 'State 1,' both 'States 1 and 2,' and 'States 1, 2, and 3 –that is, all states, or combinations, not included in the notions of *known* that the dissertation acknowledges.

While each such notion has merit, the most obvious indication of un*known* is occupancy of VSAT State 1, the state test takers themselves associated with zero target word familiarity. The sums of words in this State for each child, and for each set of texts, appear in Appendix 4 and derive from the raw data in Appendix 2a. Table 5.11 presents the results of binomial sign tests on children's zero scores for each of the 10 possible pairs of sets, with *known* words restricted to those with VSAT State 1 designations. Findings from these tests indicate two pairs of sets of texts in which the totals of unknown words proved to be significantly different at $p<0.05$: (1) Sets 5 and 1 ($p=0.004$), and (2.) Sets 1 and 4 ($p=0.013$).

<u>Conclusion 67</u>: The gains in spaced learning afforded by Set 5 express themselves in both (a) significantly more ($p<0.05$) *known* target words and (b) a statistically significant difference in the sums of words *unknown* to participants (i.e. words in VSAT State 1).

In the light of these findings, the sign testing procedures that informed the Part 1 discussion and the many indications of non-significant differences, do not quite rule out a spaced learning advantage or despite pairwise testing indicating otherwise. It is possible, in theory, for a non-significant difference in *known* word sums from reading a set of a pair to nevertheless co-exist with a significant difference ($p<0.05$) in the number of children learning no words from reading those same two sets. Unsurprisingly, the 'zero learning' observations indeed suggest certain sets of texts as somewhat more learning conducive than others –namely, Sets 4 and 5, compared to Set 1. These two pairings stand as supplementary, qualifying, evidence to the Part 1 observation of so few instances in which the participant group gained more words from relatively distributed learning opportunities. The zero score data, in short, implies a spaced learning advantage that evades detection under the two testing procedures the current study employs.

	Set1 Set2	Set 1 Set 3	Set 1 Set 4	Set 1 Set 5	Set 2 Set 3	Set 2 Set 4	Set 2 Set 5	Set 3 Set 4	Set 3 Set 5	Set 4 Set 5
sig. (2-tailed)	.503	.143	**.013**	**.004**	.791	.263	.092	.549	.227	1.000

Table 5.11: Binomial sign test findings; sums of target words in VSAT State 1.

5.25 Research Question 1: Conclusion

Research Question 1 asked:

> *How significant (statistically and pedagogically) are the differences in sums of novel words child readers gain from encountering those words under more, or less, distributed learning conditions? (Does, for example, more learning arise from 'x' reencounters with word 'y' during a single daily RR session than 'x' reencounters with the same word over several daily sessions?).*

The data gathered during the present study leaves little doubt that how often Year 4 Thai first language children encounter a novel word during RR sessions impacts upon their vocabulary gains, albeit in the restricted sense of learning word/meaning associations. Furthermore, the study finds that differences in the sums of words learned could prove statistically significant at the conventional alpha value of ($p<0.05$) despite a binomial sign test based methodology biased towards null hypothesis preservation (Section 4.21.1). The study, however, derives few generally applicable conclusions in regard to spaced learning efficacy, noting the likelihood of statistically significant gains depends critically upon the combination of lexical competencies –expressed in VSAT States– that one accepts as indicative of word *knowing*. Under the most demanding test acknowledged, a conception of '*known*' that limits *known* words to those for which a child could supply in a syntactically and semantically well-formed clause (i.e. the word occupied VSAT state 6) the study finds that readers failed to gain statistically significantly more target vocabulary however spaced the target word presentations. The evidence proved unambiguous, results from a Friedman's ANOVA and 'pairwise' binomial sign tests missing significance by comfortable margins (the nearest to a significant result emerged from children having read Set 5 as opposed to Set 3; $p= 0.167$). Indeed, for 6 pairs of sets from the 10 possibilities, p-values amounted to $p=1.000$.

Under a more expansive test (occupancy of States 6 and 5) that includes within the *known* category any words for which a child could supply a native-like synonym, the 'pairwise' test procedure identified 4 cases where the participant group gained statistically ($p<0.05$) more vocabulary from a set of a pair (see Section 5.8.2), p-values ranging from a low of $p=0.01$ (Sets 5 and 1; 5 and 2; and 5 and 3) to a high of $p= 0.03$ (Sets 4 and 2). The same 'pairwise' testing likewise identified statistically significant differences in the *known* word sums under the third, and broadest, conception of *known* (i.e. word occupancy of States 6+5+4), a definition that includes within the *known* category any words for which children could supply a loose synonym (Section 5.9).

For the two notions of *known* (i.e. VSAT States 6+5, or States 6+5+4) associated with more learning from manipulating target word presentation time, the additional hours over which a set of texts need present its embedded non-words for the participant group to gain statistically more words from reading one set of a possible pair rarely proved the same, or even comparable. For example, the statistical difference in the sum of *known* words from reading Sets 5 and 3, with *known* denoting word occupancy of States 6+5 fails to emerge when the *known* category includes target vocabulary occupying States 6+5+4. Even under the same test of *knowing*, the time interval between word reencounters that sufficed to ensure significantly more word gains from reading one set of a pair might not suffice in the case of an alternative pairing: The 72-hour difference in presentation time that 'apparently' proved sufficient for the participant group to gain significantly more words from reading Set 5 as opposed to Set 2 (under both the 6+5 and 6+5+4 tests of *knowing*)[108] did not ensure significantly more learned words from having read Set 4 as opposed to Set 1.

For pairs of sets of texts for which 'pairwise' testing identified significant differences in *known* word sums (Section 5.10), the study reveals the source of differences as either: (1.) a preponderance of words occupying a particular VSAT state from among those defining the sense of word *knowing* of interest (that is, State 6 or 5 given the 6+5 definition of *knowing*, or States 6 or 5 or 4 under the 6+5+4 definition), or (2.) from additional words dispersed rather more evenly among those same states. In the latter case, the sum of words in State 'x' from having read the texts of a set did not differ significantly from the sum occupying that same state from reading those of the other ($p<0.05$). In the few cases where the participant group gained significantly more words to a particular VSAT state standard, that state amounted to State 5 in all instances. The study argues that this state denotes a low-level productive competence, insufficient for native-like productive word use yet nonetheless of some practical communicative value. Purposes for which students might find State 5 familiarity helpful in classroom settings might include, among others, simple expressions of wants, understanding written texts, and performing satisfactorily on multiple-choice vocabulary tests. Clearly, the additional State 5 words attributable to a spaced learning advantage could comprise a substantial proportion of the extra words from the set providing the more spaced learning opportunity. With *known* words as those in States 6+5 the additional State 5 words[109] from reading Set 5 over those from reading Set 1 amounted to 80 percent of the extra (i.e. the sum of nouns, verbs, adjectives and adverbs) gained from a spaced learning effect (see Section 5.8.3). From reading Sets 5 and 2, the sums in State 5 amounted to an only marginally less impressive 72.7%. An unanticipated finding from the study were instances where the participant group gained more words to a particular VSAT competence from a set of pair, despite failing to gain significantly ($p<0.05$) more overall (i.e. the total *known*, undifferentiated by lexical

[108] This example holds true under either the 6+5 or 6+5+4 test of *knowing*.
[109] In those cases where the sums of words in this state were significantly different.

class) from reading that same set (see Section 5.8.3). An example is the non-significant difference in the excess State 5 words from reading Set 4 over those from Set 1, with *known* words defined as those in States 6+5. This finding would have escaped notice had the research failed to differentiate notions of *knowing* into various lexical sub competencies (States).

An issue arising in any discussion of significance concerns just what learning amounts to in terms familiar to teachers, linguists and others interested in pedagogical efficacy and classroom practice. The data from VSAT testing reveals that a spaced learning advantage expresses itself in numerically small, incremental, word gains that only become substantial from regular, voluminous, reading over relatively long time spans. Nagy et al. (1987) stress the point in their study of English L1 children's vocabulary uptake noting that it holds true irrespective of grade level and, quite likely, across genres. The present investigation extends the generality of this observation to young, Thai, EAL primary children attending international schooling. The repeated rounds of VSAT testing with the same children, coupled with learning failures (especially under the State 6 notion of *known* word), implies that 'meaningful' vocabulary development from RR presupposes a well-established reading habit along with regular opportunities for pupils to engage recreationally with interesting texts. This said, a regular diet of reading materials apparently well-suited for learning unfamiliar words –the Set 5 texts stand out as the most obvious example– did not invariably see children making impressive vocabulary gains in the short term (i.e. a few days or weeks). The learning recorded from the current research, as elicited through VSAT testing, arose from multiple exposures to the same word in researcher-prepared contexts from which a reader might deduce its meaning. Even after such apparent learning as the participant group displayed, it remains unknown for how long any gains persisted beyond the testing session. That a child would successfully recall a target word a month or two after completing a set of texts seems doubtful in the light of investigations into memory decay (Waring & Takaki, 2003). Which sets of texts best optimize spaced learning opportunities with vocabulary expansion in mind, and given the school's timetabling, remains far from obvious, however, since any assessment calls for evaluating totals of words learned together with the child's knowledge depth of those same words. The teacher setting out to adapt RR materials with vocabulary expansion in mind would observe somewhat different totals, and quality of word gains, depending upon which set of texts among those the current study employs served as the model for the adaptations. While Set 5 indeed seems particularly conducive to promoting overall gains, Sets 1 and 2 emerged as noticeably less so. Even this observation, however, may not hold up in regard to words assigned to VSAT State 6 given that from no set of texts did the participant group gain statistically more words than from reading any other. However worthwhile designing texts to promote word learning might seem, the practical challenges of textual adaptation mean that such efforts will not suffice in themselves. As in the current study, word learning gains stem from both 'optimized' texts but also teachers able and willing to successfully integrate those texts into an RR program having paid due regard to factors such as, notably, the number of reading sessions per week and their duration.

To examine pedagogical significance (the second limb of Research Question 1) the study draws upon projections of learning gains from reading one experimental set as opposed to another, employing a standard measure construct to evaluate learning outcomes. Over the course of a single academic year, a period that traditionally defines the duration of school-based programs, the study finds children could potentially gain useful additional words from texts offering more spaced learning opportunities. These gains ranged from 482 extra words from reading Set 5 as opposed to Set 1[110] to a low of 139 from reading Set 4 as opposed to Set 2.[111] In standard measure units, the participant group's extra learning (speculatively) amounted to between 69% (Sets 5 and 1; 0.69 SM) and 20% (Sets 4 and 2; 0.20 SM) of words from reading authentic materials over the one academic year period (See Fig. 5.38, p. 233). Expressed as time savings, these figures imply up to 118 additional RR sessions per academic year for a child to learn as many words from texts less optimized to exploit spaced learning than those more so. Even for the pair of Sets 4 and 2, the pair associated with the lowest SM gain,

---

[110] This assumes *known* words as those in VSAT States 6+5+4.
[111] This assumes *known* words as those in VSAT States 6+5.

children would require 53.7 extra RR sessions with Set 2 type texts beyond the yearly 148 total to learn the total words they would from reading sets designed to the specifications of Set 4 (see Fig. 5.39, p. 235). Certain caveats apply, however. Claims that spaced learning opportunities contribute to vocabulary expansion mean little without specifying what exactly amounts to a *known* word, an issue upon which not all teachers or linguists agree. While the participant group gained significantly more words from set 5 than Set 3 under the State 6+5 test of *knowing*, they did not do so under the equally acceptable (to some) alternative 6+5+4 test (Section 5.18). For the two notions of *known word* associated with significant differences in learning outcomes (i.e. word occupancy in either VSAT States 6+5, or 6+5+4), only from 4 pairs of sets of a possible 20 did a significant difference in learned word totals arise. Bearing in mind that from no possible pairing of sets did a significant difference in target word gains emerge under the State 6 test of *knowing*, this becomes rather less impressive still: i.e. a mere 4 out of 30 cases, or just 13%.

The reports of learning outcomes of individual students (Part 1) make it possible to compare learning between the participant group as a whole, and single members of that group. Perhaps most obviously, we see that for any set of texts (of a pair) from which children collectively gained significantly more words, several participants nevertheless gained the greater sum from the other. In no cases, however, do we find a pattern in which a child consistently learned an equal, or larger, number of words from the set affording the more massed target word presentations –i.e. more gains from Set 1 than from Set 2, larger gains from Set 2 than Set 3, etc. Despite the several unambiguous cases of a spaced learning benefit at the group level, the study likewise finds few children displayed the opposite outcome –i.e. an invariant (without exception) pattern of additional words from the set (of a pair) offering the relatively distributed word presentations; this remains true irrespective of how expansively (or otherwise) one defines a *known* word. Under the most inclusive test of *knowing* (occupancy of States 6+5+4) only 8 participants invariably gained an equal, or additional, sum of words from the set of a pair providing the more distributed target word encounters; fewer still (just 4) displayed a pattern of more or equal gains given the States 6+5 definition of word *knowing,* and just 6 children once we restrict *known* words exclusively to those that occupy State 6. The variability in learning successes between children rules out reliable long-term projections of word gains for individual students from group performance data. A participant could gain as few as 1 target word from a set of texts, as did student 18 from reading Set 5 (student 18),[112] only for another to successfully learn all four (student 5). The data in Appendix 2 contains many other such examples.

Prior to the study, it remained unclear which pairing of sets would associate with the greatest disparity in *known* word sums i.e. from which combination of sets the largest difference in learning would arise. The investigation leaves little doubt, however, that the massed encounters with novel words from reading Set 1 ensured far fewer word gains than did the relatively spaced encounters afforded by Set 5. This is not, however, to deny a possibly massed learning advantage should the time intervals between meeting the same word prove unduly excessive (See, for example, Section 5.8.1 that discusses sums of words in State 6). One can imagine texts, for example, in which an excessive interval between meeting the same word ensured no reader recollection of the word when meeting it again later in the script. Nor does the study discount possibly larger differences in vocabulary gains had students encountered novel words under rather more distributed conditions than Set 5 afforded (i.e. 120 hours). The optimum time interval between meetings with the same unknown word for maximizing vocabulary learning from regular, long-term, RR opportunities, remains unknown; the issue lies beyond the narrow scope of the dissertation's Research Questions.

Despite Set 5 type texts proving relatively learning conducive, the present research offers no assurance such texts capture the ideal interval between target word presentations, vocabulary development foremost in mind; the findings support the more modest claim that for participants in the present investigation, and given the design of the reading materials they engaged with, vocabulary gains from Set 5 type texts exceeded those from texts designed to the specifications of Sets 1 and 2. The conclusions hold for the circumstances in which RR took place in the host institution (i.e. four

[112] The example assumes *known* words are those in States 6+5 (see Appendix 2b).

sessions per week, each session of around 25-30 minutes duration) and given the demographics of the participant body (see Section 6.4). Whether findings apply to somewhat broader, more diverse, populations remains unestablished.

5.26 Research Question 2: Conclusion

Research Question 2 asked:

> *How significant (statistically and pedagogically) are differences in the sums of novel words of the four content word classes (nouns, verbs, adjectives and adverbs) child readers gain from encountering those words of a particular class under more, or less, distributed learning conditions? (Does, for example, more learning arise from 'x' reencounters with noun 'y' during a single daily RR session, than 'x' reencounters with noun 'y' over several daily sessions?).*

The Question takes us beyond aggregate measures of word gains to explore findings at the level of lexical class. It aims to determine the likelihood, or otherwise, that distributed learning accounts for relatively more gains of one content word type than another and how alternative notions of word *knowing* may moderate observed differences in learning successes.

Drawing upon McNemar test results, the study found that spaced learning reasonably explains statistically significant differences in children's noun and verb gains from having read certain texts as opposed to others. No such significant differences, however, emerged for either adjectives or adverbs, irrespective of any possible pairing of sets and/or notion of *known* word that the study acknowledges. The evidence for relatively more noun and verb learning from spaced target word presentation often proved compelling (see Sections 5.12 and 513); likewise, findings that rejected any such association could prove equally unequivocal. Regarding adjectives, for example, under the 6+5+4 test of *knowing* the nearest to a significant difference in *known* sums came to an 'unimpressive' p=0.227 (McNemar) from the respective totals the participant group gained from Sets 1 and 2. Under the 6+5 test of *known* word the closest to a significant difference is p=0.180 (Sets 3 and 4), and under the State 6 test of *knowing*, p=0.388 (this for the pairs of Sets 2 and 3, and 3 and 5). Similarly high (non-significant) p-values arise when comparing adverb gains. The closest to a significant difference for words of this class emerges from children having read Sets 2 and 5 with *known* words restricted to occupants of States 6+5+4 (p=0.263), a figure strongly consistent with the null (no effect) hypothesis. Under the States 6+5 test of *known,* the nearest to a significant difference arises from children having read Sets 1 and 3, but still amounts to an 'unimpressive,' p=0.388. Limiting *known* adverbs to those in State 6, the closest to a significant finding emerged from the difference in learned word sums from reading Sets 4 and 5 (p=0.626).

Just as gross *known* word totals varied depending upon the test of *knowing* applied*,* so too the sums of words of a single lexical category. The current study identified several cases of the participant group learning statistically more ($p<0.05$) nouns or verbs under one definition of *known* and yet failing to do so under another. For example, children gained a significant addition to nouns from reading Set 5 over those from Set 3, assuming *known* denotes word occupancy of States 6 and 5, though conspicuously failed to do so under the more expansive notion that recognizes *known* words as any occupying either VSAT States 6, 5 and 4. If indeed each notion of *known* the study employed appears objectively defensible it follows that those involved in children's education may hold very different, and yet equally valid, opinions regarding just how substantially spaced learning contributes to children's learning outcomes from RR sessions.

That children did not gain significantly more of any single class of words from a set of texts of a possible pair, with *known* words as those meriting VSAT State 6 designations, would most concern teachers whose interest in RR lies in its potential for promoting gains of the productive knowledge that underlies writing and speaking; as we have seen, the insensitivity of sums of words children could incorporate into 'syntactically' and 'semantically' well-formed clauses reveals itself in both the 'pairwise' and 'general' testing procedures. For others –teachers rather more interested in fostering receptive word knowledge skills– however, adapting texts to exploit spaced learning might seem a rather more reasonable proposition. Even so, the study finds that children's word gains proved variable in quality (State 4 seems resistant to a spacing effect, for example), and not always numerically impressive, predictable, or even unambiguously attributable to a spaced learning advantage. The additional 10 nouns from reading Set 5 over Set 2 (*known* words as those in States 6+5) primarily occupy State 5, for example, though attributing these gains to spacing proves contentious given a p-value from McNemar testing of respective State 5 sums exceeds (albeit marginally) the conventional $p<0.05$ –see Section 5.12.3. In contrast stand the pair of Sets 5 and 3, where significantly more nouns from having read Set 5 arises from a general increase in the totals of nouns in states 5 and 4, as opposed to a significant increase in the number occupying either one state or the other; indeed, the respective State 5 *known* noun sums from having read Sets 5 and 3 miss significance even at the relatively uncommon alpha = 0.1. The cases illustrate the pedagogical point that for the same sum of additional verbs or nouns arising from a more spaced learning opportunity, a reader will not necessarily *know* those extra words (of the class of interest) to the same standards of 'depth' in each instance –i.e. the proportion of words in States 5 and 4 could vary, and often quite markedly so. A particular distribution between these states that strikes one teacher as impressive might seem less so to another, depending upon the value s/he happens to attach to productive as opposed to receptive word knowledge skills.

To explore pedagogical significance, the study adopted the same notion of standard measures (SM) applied when examining gross *known* sums, except that a standard measure now denotes the sum of words of a *particular* lexical class gained from unadapted texts over an academic year period. The robustness of conclusions from the analysis depends upon two factors underlying the computation of the SM values themselves: (1.) The accuracy of the Nagy et al. (1987) estimation of the total words a child gains from reading unadapted texts, per academic year, and (2.) the definition of *known* word one chooses to acknowledge. With *known* words as those in VSAT States 6+5, children's noun gains from Set 5 over those from Set 2 amounted to 0.38 standard measures, and a rather more impressive 0.61 SMs if *known* words constitute any occupying States 6+5+4 (Section 5.20). Expressed as additions to learning time and/or RR lessons, the study predicts a child might require between 69.7 and 119 additional RR sessions, and between 29.04 and 49.1 extra hours of reading to learn the equivalent number of words from the set of texts (of a pair) presenting target words under less optimally spaced learning conditions.

The differences in *known* word sums from reading the experimental sets on the one hand, and unadapted on the other, leave little doubt that textual adaption selectively impacts upon the total words children gain of particular lexical classes. However, while sets offering more spaced target word encounters indeed saw children gain significantly more ($p<0.05$) nouns and verbs –albeit under particular configurations of circumstances this may offer little comfort to those hoping for across the board gains in all four content word classes. For such teachers, and others, the non-likelihood of more adjectives or adverbs from any single set of a pair suggests that textual adaption may have little to recommend it; in other words, the design specifications of any one set of texts would likely prove no more useful a blueprint for adapting authentic reading materials than any other. For nouns and verbs, presumed gains proved variable relative to those from authentic reading materials over the hypothesized one academic year period. Under the 6+5 definition of *knowing,* children would, it seems, gain fewer nouns and verbs (Section 5.22) from texts designed to the specifications of Set 5 (the most optimal for learning) than from authentic texts from their RR sessions –an observation the more surprising given that each set of texts incorporates features to facilitate vocabulary uptake (Section 4.8). The factors potentially explaining the authentic texts' apparent advantage still remain largely unclear. Tentatively, the explanation lies in the VSAT State 6+5 notion of *knowing* amounting to a somewhat more stringent test of *known* than Nagy et al. (1987) adopted in their estimation of the chance of word learning from a single exposure –a test that the current study employed to derive the unadapted yearly learned word total (see p.228-230). As noted, less impressive gains in SM units will arise should the Nagy et al. test of *known* prove easier to satisfy than a VSAT alternative, just as higher SMs emerge if achieving a *known* designation under Nagy et al.'s (1987) test prove relatively more difficult (p.236). How similar are either of the two 'VSAT based' conceptions of *known word* (namely the State 6+5, and State 6+5+4 based definitions) to the notion Nagy et al. (1987) conception remains unclear.

5.27 Conclusion

That children gain word meanings from reading might seem little short of stating the obvious. However, widely disseminated, research-validated, demonstrations of such learning only really emerged in the 1970s, the first major study of incidental vocabulary development arguably that of Saragi, Nation and Meister (1978) who investigated adult EAL learners' uptake of non-words in Burgess's novel *The Clockwork Orange*. Subsequent research has helped clarify both the circumstances in which gains arise, and the particular populations for whom reading reliably predicts vocabulary expansion. The current study contributes to this ongoing research endeavor by demonstrating reading induced vocabulary development among young, Thai, EAL learners attending a Bangkok based international school; it appears to be the first such study to do so, and one of only a handful of research projects to report upon word gains among primary aged children (the most notable among other such studies are Nagy et al., 1985 and 1987). Students participating in the present research project unambiguously learned novel word meanings under spaced and massed learning conditions, irrespective of what we accept as a *known* word, and from texts which, in the researcher's view, engaged children as much as did their regular RR reading materials. Of the total 28 participants, each learned at least one word having read all five sets of texts, with many children gaining substantially more. Nineteen children (or 68% percent) learned one or more words under the most restrictive definition of *known* (word occupancy of State 6) and from the set associated with the largest number of non-scorers (Set 2). For Set 5, with *known* words defined as those in States 6+5+4, every child gained two words or more.

The study claims pedagogical relevance in that all sets of texts are designed to specifications that teachers might reasonably acknowledge as helpful for maximizing word learning gains (Section 4.8) during RR sessions. Every set, for example, contained an optimum density of unknown words (less than 2 percent); each embedded a wealth of contextual clues hinting at the meaning of the *unknown* terms; the texts made exclusive use of highly decodable vocabulary and each proved sufficiently similar to authentic reading materials that the scripts passed as such among one or more teacher assessors (Section 4.17). Drawing upon target word gains, the study demonstrates conclusively that learning successes could, albeit not invariably, depend much upon the time intervals between target word encounters during reading experiences. For teachers and others interested in adapting texts to promote receptive vocabulary gain, and especially if prioritizing gains for the particular lexical classes of nouns and verbs, the contribution of more or less spaced presentations to learning outcomes deserves attention and recommends itself for future research. The study affirms that spaced learning impacts upon how many words children learn from recreational reading and the depth of lexical understanding they gain of any new vocabulary items.

Chapter 6

Implications and Conclusion

6.1 Introduction

Rich vocabularies allow children to interact more effectively with their peers, access school curriculums and perform satisfactorily in ever important public exams. We now know that vocabulary knowledge predicts reading comprehension (Anderson & Freebody, 1981), that poor vocabularies explain failures to meet academic norms (Baumann, Kameenui, & Ash, 2003; Becker, 1977), and that lexical insufficiencies both foster disinterest in reading and detract from the quality of children's written compositions (Laufer & Nation, 1995). For such reasons, and many others besides (see, for example, Nation, 2000), understanding the causes of vocabulary deficits and possible means to remediation has long attracted academic interest and remains a popular research topic among linguists and teachers alike. The current chapter explores, and highlights, a few of the theoretical and practical contributions of the present study towards informing vocabulary and literacy instruction with EAL students needs foremost in mind. It concludes with a reference to the host institution and the relevance of findings to international schools generally.

6.2 Theoretical implications of the current investigation

The dissertation contributes to ongoing discussion of three topical issues in particular:

1. The effects of overabundant learning opportunities in the context of attention-raising devices,
2. The relative learnability of words depending upon lexical class, and
3. The contribution of spaced learning to a comprehensive, robust model of word-meaning gains.

6.2.1 Overabundant learning opportunities and attention-raising devices

A child gains new word meanings from a conscious learning effort (Section 2.2), implying that textual devices aimed at inducing a learner to *notice* an unfamiliar word will raise the likelihood that s/he successfully forms a match between an orthographic representation and the concept it denotes. The noticing hypothesis (Schmidt, 1995, 2000), the basis for this claim (Section 2.2), has prompted investigations into various textual enhancements such as bolding, italicization and marginal glossing as means to promote reader awareness of novel words and/or particular target grammatical structures (see, for example, Hulstijn, Hollander, & Greidanus, 1996; Jenpattarakul, 2012). A substantial body of literature validates such augmentation devices for raising a target feature's salience (Nation 2001). Predictably, perhaps, enhancements of various kinds have now found their way into children's texts, and especially so works of non-fiction (see Dufon & Hong 1994 for a discussion of glossing and word retention).

Assuming the present study's conclusions extend to declarative learning in general, textual enhancements will likely contribute more, or less, to a child's sum of learned knowledge according to how often, and when, the enhancing device prompts *noticing* of the particular feature it references. The actual time interval between such *noticing* events would depend upon the child's reading speed and the duration and frequency of reading opportunities –the same factors, that is, that controlled when children encountered novel words in experimental texts during the current investigation. The mechanics of how the time intervals between *noticing* opportunities promote learning (whether of vocabulary or other textual features) still remains unclear. Writing of vocabulary, Nation (2000) claimed short intervals between word reencounters would discourage the deep, 'learning facilitative,' cognitive processing that arises when readers experience recall difficulties. That children in the present study made relatively unimpressive vocabulary gains from reading Set 1 as opposed to Set 5 texts, in this view, stems not from insufficient *noticing* opportunities as such, but rather the timing of those opportunities during RR sessions. Likewise, one might reasonably suppose participants learned less from Sets 1 and 2 than Sets 4 and 5 because the relatively short interval between meeting the same word ensured a memory trace persisted from the time of prior encounter. Hintzman (1976) termed this association of processing effort, trace decay and learning, the Deficient Processing Hypothesis, arguing that it stood as a plausible alternative to encoding variability theories of spacing effect expression. The 'all-important' interval duration depends most obviously upon the number of intervening words between one occurrence of the device (or novel word) and its next appearance, along with the child's reading speed in words per minute. The more the intervening words, and the slower the reading, the longer the interval between the next encounter with the same textual enhancement or unlearned vocabulary item.

This still leaves open the question of just what amounts to the optimal time interval between reencounters for promoting learning gains. According to Bahrick (1984), most learning arises when a reader meets a textual feature (be it an unknown word or otherwise) at such time as previously gained knowledge of that feature remains only barely recallable. This implies short intervals initially, leading to progressively longer as a memory trace becomes more established in the long-term memory store. The current study sheds no light on possibly more impressive word gains had the interval between a particular target word's occurrence increased by some factor. Even so, the common learning processes underlying 'word gains,' and declarative knowledge more generally (e.g. dates of events, names, chemical formulas, to cite just a few) suggest that manipulating the time intervals between encounters (i.e. degree of spaced learning) may comparably affect gains of each. Assuming, with Bloom (2000), that the same learning process sustains word meaning gains and that of 'factual explicit knowledge,' then textual enhancements will likely promote learning most effectively if memory decay of the feature they reference evokes an effort to recall that same feature when the device next brings it to the reader's attention (See 2.6.1). Whether a particularly conducive interval for gaining word/meaning associations proves equally so for mastering other language features or details that an enhancing device raises to attention, calls for further research. Drawing upon word gains from experimental texts, the current study suggests that optimally spaced repetitions of the same textual enhancement should positively impact upon learning of declarative facts generally.

6.2.2 Word class

In *Learning Vocabulary in Another Language*, Nation (2001) discusses the “intuitively” plausible notion that words of certain lexical classes prove more readily learned than others. He cites nouns as the oft-presumed most gainable class, followed by verbs, and then adjectives and adverbs in that order. Nation (2000) goes on to argue that evidence for this hierarchy nevertheless remains equivocal, agreeing with Laufer (1997) that “commonsense” presumptions often lack credible research validation. Investigating relative learnability has, however, proved difficult given the challenge of isolating the contribution of lexical class from other factors potentially affecting learning gains. Laufer (1997), for example, singles out a particular concern arising with the notion of concreteness –a measure of the perceptual salience that an entity/state denotes (compare *water* with *courage*). Linguists have long argued that the more concrete the denoted entity/state a term refers to, the generally easier learning and retention becomes (Nation, 2000). Where concreteness proves controversial is in the role it may play in explaining a much hypothesized noun-learning advantage (Schwanenflugel, 1991). To validate that nouns indeed prove easier to learn than other word classes raises a critical question: Can one argue that nouns prove easier to learn as such (i.e. because they belong to the lexical class of noun, and all this entails in terms of affixation, distribution and selection restrictions) or does learning ease arise from their ready ‘visualization,’ implying a possible learning advantage associated with issues of physical perception or imageability? With concreteness a potentially confounding variable, selecting suitable target vocabulary for learnability research becomes critical, the investigator having to control for concreteness ‘across’ word types (i.e. among nouns, verbs, adjectives and adverbs, for example), but also ‘within’ the class of nouns themselves; some nouns will, after all, exhibit substantially more concreteness than others (compare ‘air’ with ‘brick,’). In the past, investigation into the learnability issue have yielded contradictory findings depending upon the choice of words from within the lexical classes the researcher has chosen to compare (Schwanenflugel, 1991).

The current study makes two contributions to the learnability debate, one methodological, the other substantive. Methodologically, the study identifies the time intervals between word reencounters along with notion of *known* word as additional confounding factors that learnability research must necessarily acknowledge. The evidence appears unambiguous and compelling. From reading experimental Set 1, for example, and with *known* words restricted to those in VSAT State 6, sums of *known* nouns and adverbs proved significantly different (McNemar, p=0.003), as did sums of nouns and verbs (McNemar, p=0.013). From reading Set 3, likewise, a Cochran's Q test revealed one or more significant differences in the proportions of words gains (Q=13, df=3, p=0.005); McNemar follow up tests subsequently narrowed this down to a discrepancy in adjective and adverb sums (p=0.031). Under the 6+5+4 test of *knowing*, the participant group's total for *known* nouns from having read Set 5 significantly (p<0.05) exceeded that for adjectives (McNemar, p=0.031) despite that sums for these same word classes from having read Set 3 miss significance by a comfortable margin (McNemar, p=0.180).

The study's substantive contribution amounts to the light shed on the hypothesized noun learning advantage. Participants clearly learned more nouns than words of other classes, albeit alternative understandings of *known*, together with target word presentation time moderated learning outcomes. From Set 5, for example, and with *known* words as those in States 6+5+4, children's noun gains exceeded those for adverbs by 11, a highly significant difference at p<0.001 (McNemar). Nouns likewise proved relatively easier to learn than adverbs from reading experiences with Set 3 (p=0.039), as noted. The evidence for relative learning ease does not, moreover, arise solely under the most liberal test of *knowing* the study examined. Given the relatively massed presentation associated with Set 2, children gained statistically significantly more nouns than adverbs (McNemar; p=0.022), with *known* words defined as those occupying VSAT States 6+5, while failing to gain significantly more nouns than either verbs (p=0.18) or adjectives (McNemar; p=0.503). The percentage differences in *known* noun sums, compared to gains among other word classes, could prove relatively large. From reading Set 5, *known* verbs (n=6) amounted to just 40% of the total *known* nouns (n=15), if *known* denotes word occupancy of VSAT State 6. Noun learning ease, contrasts most strikingly with children's difficulty gaining adverbs. Under a test that limits *known* words to those in States 6 and 5 children gained just 10 adverbs compared to 27 nouns (Section 5.5.1) –an almost threefold difference.

6.2.3 The overall model

Directly, or indirectly, most vocabulary research aims to contribute towards a robust model of word gains from reading experiences that successfully incorporates the totality of factors accounting for learning outcomes. Many of these critical factors are already familiar and the subject of extensive prior research; examples include the density of a text's unfamiliar lexis (Laufer, 1992), reader interest in the script (Krashen, 2004), the reader's decoding skills (Gough & Tumner, 1986), the helpfulness of contextual clues (Ames, 1966), the ability to apply strategies to recover word meanings, and the background knowledge a reader brings to the reading task (see Nation, 2001 for a review). Historically, however, model building has always tended to focus upon quantifying the independent contribution of factors to learning outcomes, and the effects of plausible moderating variables. The task of synthesizing these factors into an explanatory account of word learning has progressed rather more slowly. Partly this stems from alternative perspectives as to how reading proceeds (see Snowling & Hulme, 2008 for a discussion), but also from uncertainties concerning the conjunctive contributions of relevant factors, as opposed to their individual impacts.

The concern of this dissertation lies squarely within the first research tradition –the effort to seek out factors contributing to word-meaning derivation and subsequent storage in long term memory. While the study firmly establishes spaced learning as an additional factor, the narrow research focus still leaves many questions unanswered: Could, for example, a spaced learning advantage prove less evident depending upon textual genre? What moderating effect does background topic knowledge have on the efficacy of spacing target word encounters? Under what circumstances does spaced learning give rise to more, or less, learning relative to other factors potentially influencing learning outcomes? Despite these limitations, the investigation assists model building both *directly*, and *indirectly*. *Directly*, the study presents a research-validated claim that viable models of vocabulary learning (Chomsky, 1956) must indeed acknowledge the time intervals between reencountering the same novel word as contributing to the 'depth' and 'breadth' of a child's word gains. *Indirectly*, the study highlights the critical need for standardizing crucial research terms. These include just what the label '*known* word' denotes, and what *exactly* we mean by receptive and productive competence. For pedagogical purposes, common notions of what typifies native speaker children's lexical knowledge at different grade levels seems helpful for EAL and mainstream teachers alike. The search for rigorously defined key terms recommends itself as a sensible topic for further investigation.

6.3 Practical implications of the current study

The study has several implications for teaching practice, both for those engaged in vocabulary instruction generally (EAL teachers in particular) and for the staff of the host institution in which the research was conducted. The current section considers three such implications, the choice informed by several the topical concerns of those engaged in vocabulary research and/or teaching in international education:

1. The number of encounters with a word that suffice for its uptake
2. The effect of learner ability on word (meaning) learning
3. The effect of spaced learning on gains in receptive as opposed to productive word use

6.3.1 The critical number of word encounters

A popular subject of incidental word-learning research is the number of encounters with an unfamiliar that suffices to establish its meaning in long-term memory. Findings have differed across studies depending upon choice of methodology, the researcher's theoretical perspective, the data-elicitation instrument employed and the demographics of the participant population. Webb (2008) cites several commonly quoted claims:

> *Rott (1999) suggested that six encounters may be enough to learn a word. Horst, Cobb, and Meara (1998) suggested eight encounters are needed, Saragi, Nation, and Meister (1978)*

> *suggested 10 encounters, Webb (2007) suggested that more than 10 encounters are needed, and Waring and Takaki (2003) reported that it may take more than 20 encounters to incidentally learn the meaning of a word* (para.1).

Findings from the current study neither endorse one claim or another, arguing, rather, that differences in these figures arise partly from alternative conceptions of what word *knowing* might reasonably imply. Holding the number of target word reencounters constant (12 in the case of this dissertation), for example, the probability a child gains sufficient familiarity with a novel word to permit productive use in syntactically and semantically well-formed clauses (the strictest test of *knowing* the current study recognizes) ranged from 0.21% to 0.28%. The figure rises to between 0.44% and 0.63% under a less restrictive notion that extends the class of *known* words to those a child could define in native-like terms, and between 0.62% and 0.82% under a liberal test that accepts as a *known* word any that VSAT testing assigned to States 6,5 or 4.[113] Depending upon the test of *known* that one acknowledges, each learning/encounter estimation Webb (2008) cites becomes more, or less, plausible. No single definition of *known*, however, stands out as objectively 'better' than another. A word a child *knows* to the standard of VSAT State 5 may strike one instructor as truly warranting a *known* designation but not necessarily a colleague for whom a State 6 lexical competence signifies the reasonable minimum standard. Disagreement might most likely arise in regard to vocabulary occupying State 4: Can one reasonably assume a child *knows* a word if, as in this case, s/he lacks the capacity to use it productively and evidences only a rudimentary standard of receptive competence? Alternative notions of *known*, and their respective merits, is no less relevant an issue when evaluating probabilistic estimates of word learning from a single meeting (see, e.g., Horst, 2001; Nagy et al., 1987). A 5% probability of gaining a word from a single textual encounter means little without specifying what that gain amounts to in lexical understanding.

---

[113] Note: These are the probabilities associated with 12 encounters with the same word.

A second difficulty with 'number of encounters' estimates relates to lexical category. The current study finds that likelihoods of word gain depend upon which of the four content classes the word belongs to, given notion of word *knowing* of interest. Children participating in the present research gained significantly more nouns from reading Set 5 (given 12 encounters), for example, than they did from reading Set 2 with *known* words defined as those in VSAT States 6+5, just as they gained more verbs from reading Set 5 than Set 1 with *known* denoting occupancy of States 6+5+4 (Section 5.14); crucially, however, the participant group did not learn significantly more of either class, when restricting *known* words to those in VSAT State 6. The chance of the 'average' child learning an adjective from 12 encounters ranged from a low of 17% (Sets 2 and 5) under the State 6 test of *knowing* to a more creditable 78% (Set 5) if by *known* we include words in any of the States 6,5 or 4. For adverbs, the least learnable of lexical classes examined, the likelihood of gain from 12 encounters lay between 10% percent (State 6, Sets 1,2,3, and 4) and 39% (States, 6+5+4, Set 2). The probability of a reader's learning success depended both upon the notion of *known* word and the specific set of experimental texts s/he engaged with. From reading Set 5, and with *known* words as those in States 6+5+4, the chance of a noun gain from 12 encounters with the same word was 100%. For verbs this was 89% and for adjectives and adverbs, 78.5 and 60.7% respectively.

6.3.2 Learner ability

The centrality of conscious learning for vocabulary learning (Paradis, 2009; Ullman, 2004) means children's word gains become the more impressive to the extent a child draws upon metacognitive knowledge (Ellis, 2002). The intervals between encountering the same novel word, determine how much of that word a child subsequently recalls (given memory decay) and to what, in effect, s/he can apply that knowledge during meaning deduction. But what amounts to an optimal interval between reencounters of the same *word/fact/detail* for one child may not be so for another. Several illustrations come from the discussions in Part 1 (Chapter 5) of individual word gains, the data indicating cases in which a participant ranked higher than a peer in words learned from set 'x' but lower from reading set ‘y’ or, for the same set of texts, ranked differently according to how the study defined a *known* word. Such findings seem anomalous if participants’ metacognitive knowledge remained constant (as it reasonably did) during the short duration of the study, unless one supposes the spaced learning a set of texts provided did not prove equally learning conducive for each study participant. Plausibly, and while accepting the case for a spaced learning advantage under the particular conditions noted in Part 2, the optimum time interval between word encounters varies from child to child. This would hardly seem controversial or unorthodox. It tallies with Gardner’s (1983) notion of multiple intelligences, and sits comfortably with teacher observations that students will tend to embrace and favor different learning styles. The current study presents further affirmation of children as uniquely sensitive to the learning conditions they encounter and endowed with their own preferential learning behaviors, aptitudes and competencies. For a participant in the present investigation, learning proved most impressive for a given metacognitive competence when the time interval between novel word encounters was optimal for *that* participant.

6.3.3 Receptive and productive knowledge

While the study affirms an association between the intervals between meeting the same target word during reading and the likelihood of learning that word, for teachers the real issue concerns not the fact of spaced learning but what that learning amounts to in ‘real,’ ‘usable,’ terms. Gaining more words from one set of texts rather than another could imply anything from a truly useful supplement to one’s lexical understandings to little more than trivial embellishments of little practical importance. This becomes clearer when considering the divide between productive and receptive language capabilities. Of the six VSAT states, only State 6 denotes such familiarity with a word as permits a language user to employ it in grammatically and syntactically well-formed clauses. And yet sums of words familiar to the State 6 standard proved the most insensitive to differences in time intervals between target word reencounters. This has pedagogical implications for teachers and students alike. If language programs aim to cultivate the productive word knowledge underlying conversational exchanges, writing term papers, or participating in discussions, then manipulating the intervals between meetings with the same novel word would likely prove unproductive. This does not, however, deny productive (State 6 level) word knowledge gains from texts offering spaced learning opportunities –indeed, one or more study participants gained productive knowledge of target words irrespective of the Set of texts completed. It does mean, however, that a child would improbably gain significantly more ($p < 0.05$) words familiar to the State 6 standard from teacher efforts to manipulate the time intervals between reencounters with the same word. Actual cases of a child learning and subsequently supplying, in either writing or speech, a word outside of VSAT testing proved rare, however. Despite children having successfully mastered words to the State 6 productive standard from their RR experiences, the researcher observed only one instance of a target word appearing in a child's written work during the course of the investigation. The word was 'sol,' embedded in Set 4, the word appearing in a writing task in which children recalled a familiar story. This case aside, neither researcher nor assistants noticed target word occurrences in a child's language, spoken or written, other than during test sessions.

This does not 'mark out' spacing as irrelevant or unimportant when aiming to optimize texts for *productive* vocabulary gains, since children indeed learned more words for which they could supply a *native-like* definition –words in State 5- from spaced learning opportunities. True, State 5 signifies a lexical competence well short of sufficient for full productive use. Nevertheless, the state denotes vocabulary on the very cusp of State 6 placement. While spaced learning might appear to yield unimpressive additions, even to sums of State 5 words (Section 5.8.3), small gains in the learning likelihood of a new vocabulary item may translate into sizeable and meaningful benefits over the long term (Part 3, Chapter 5). The study adds support to an emerging consensus that regards RR as integral to a well-designed teaching program –a program that necessarily embodies an element of direct vocabulary instruction. The primary routes to developing vocabulary –reading, on the one hand, and formal instruction on the other– do not stand as a binary, 'either/or,' option from this perspective but rather complement one another. It would prove (arguably) as much a mistake to opt for one approach, as it would to ignore vocabulary deficits and assume they will resolve though unstructured everyday social interaction. Lest direct instruction might indeed seem sufficient in itself for vocabulary development, Nation (2000) cautions that teachers have long recognized the poor trade-off between instructional time and learning gains. Students participating in Beck, Perfetti, and McKeown's (1982) vocabulary program, for example, gained less than 80% of the target words (n=104), despite receiving five months' intensive tuition and a total of 75 half-hour lessons. No one seriously disputes that children accumulate new words slowly from RR (at least compared to explicit learning) –Nagy et al (1987) demonstrate as much; but, as Krashen (2004) reminds us in *The Power of Reading* (2004), RR exposes a child to low frequency (i.e., relatively rare) vocabulary in the context of an enjoyable and educationally rewarding activity. To this observation one might add that children gain a wealth of additional literacy skills from engaging with texts apart from form/meaning associations (see Section 1.0).

6.3.4 Implications for practice: The merits of textual adaptation

The current study attempted to shed light on an important pedagogical question: Would textual manipulation to optimize word gains from spaced learning prove viable given the time and effort involved? Findings on this remain inconclusive. The study revealed that differences in word gains from adapted texts appear 'small,' even allowing room for subjectivity of judgment. Of the sets of texts children worked through, a child would only gain substantially more words from any one if we assume opportunities to engage in regular and long-term RR sessions. Even with such opportunities, impressive vocabulary gains may not present, however, since those texts most 'learning conducive' for child 'x' could prove less so for child 'y' (Section 5.24). Whether adapting texts amounts to a productive exercise really comes down to a trade-off between the time and effort such adaptation involves, the anticipated gains, and the benefits children would derive from non-RR activities they might otherwise engage in. As for adaptation itself –the deletions, additions, rephrasing, word substitutions, and so on– access to good word-processing programs removes much of the inevitable tedium and effort involved. The process will likely prove laborious and time-consuming, nonetheless. The initial hurdle involves identifying suitable texts –no simple matter given copyright and access issues. The texts then need typing out on a word processor. Even a short Year 4 narrative could well require 45 minutes at the keyboard for the ‘amateur’ typist. The script now in digital form, it remains to insert target words at appropriate intervals to ensure the optimum (for word learning) ratio of 1 new word in every 50 (See Nation, 2000) while preserving the style and storyline of the original author. This editing calls for filler vocabulary to ‘pad out’ clauses, or make word deletions, the teacher all the while attempting to ensure any such alterations neither detract from reading enjoyment nor appear as clumsy additions to the script. Inevitably, textual adaptations of one form will entail others, including possible rewriting of whole paragraphs. The effort involved in script modification makes it critically important to identify suitable word candidates for textual embedding initially. This requires knowledge of the vocabulary already familiar to the intended readers, but also an appreciation of the new words children will find most helpful for meeting their social or academic needs. Instinct borne of familiarity with a year group might inform word selection, but it may prove necessary to consult word lists or conduct tests to identify suitable items.

6.3.5 Recreational Reading (RR) within the school

While the current study falls under the ‘theoretical’ label, it claims several practical outcomes. The research promoted a healthy academic curiosity among school staff. It encouraged discussion of RR as a pedagogical practice and raised the profile of vocabulary as a language competence. For those already favorably predisposed towards RR the study affirmed their faith in what remains a popular classroom practice. For the ‘doubters,’ staffroom debate on recreational reading and its place in the curriculum provided opportunities to reappraise beliefs in the light of colleagues input. Few staff members outright denied that reading led to vocabulary growth –no one argued that RR failed to deliver. The reservations, rather, concerned whether the benefits justified the time that RR necessarily entailed. Many colleagues regarded vocabulary gains from the current study as disappointing given upbeat reports in popular publications (e.g. *The Power of Reading*, 2004), and Krashen's well-attended recent presentation at Concordian International School (Bangkok). Doubts would surface in staff meetings and in informal discussions: Was RR really effective among the *particular* children attending our school? Did we have access to the *right* books? Wouldn't we need additional training? Upon one point, however, we could all agree: Children would continue to benefit from regular, explicit, vocabulary teaching. RR fell short of a panacea for resolving children's vocabulary deficits. It did, nevertheless, recommend itself as part of a broader, more comprehensive, solution that called for some measure of formal instruction.

The current research project prompted several further teacher initiatives. A successful spin-off included a teacher-designed board game to reinforce spelling of high-frequency vocabulary. The study might reasonably claim credit, also, for cultivating a sense that school-based studies rather than flown-in experts could offer viable solutions to vocabulary and literacy deficits, and perhaps our pedagogical concerns more generally. 'Outsiders' had much to share, true, but so too, surely, did our own staff and colleagues. Could it be we had ignored a reservoir of untapped teacher expertise? Above all, perhaps, that the study recruited from the current school population meant findings had a relevancy that textbook recommendations and advice from highly qualified 'outsiders' did not. It is encouraging that other members of staff are now actively investigating aspects of child learning.

Somewhat surprisingly, perhaps, the research attracted keen interest from the Thai Department. The Department had long expressed concern regarding the relatively poor Thai language skills children displayed compared to those attending state schooling –the explanation usually centered around the host institution's English-only policy and the limited curriculum time available for Thai language study. While our internal tests indicated students' spoken Thai compared favorably to that of peers in local government schools, they also revealed that reading and writing remained relatively poor. The somewhat novel notion that children might develop useful L1 literacy skills from reading struck many in the Department as interesting, if not altogether practical given timetabling difficulties. Over the course of the study, integrating RR into the Thai language program morphed from a fanciful, untested, notion to a subject worthy of serious study. The topic remains a talking point among Thai language instructors.

More generally, the research led to informal recommendations to management as well as suggestions that as a staff we collectively review several longstanding policies and practices.

1. (Recommendation): For English L2 children, the school should actively consider explicit vocabulary teaching as a supplement to regular RR sessions.

2. (Recommendation): The school should consider possible extension of RR as a means to improve Thai L1 children's decoding and comprehension of their native script.

3. (Recommendation): The school might usefully explore the practicality of extending RR to the secondary school population.

4. (For review): Whether guidelines from the UK in regard to literacy instruction pay sufficient attention to the vocabulary needs of our student population.

5. (For review): Whether the current stock of materials currently employed for RR purposes is sympathetic to the academic and vocational interests of pupils.

These various issues were, at the time of writing, currently receiving attention from the school management.

6.4 Limitations

The limitations of the current study fall into two categories: (1.) 'scope' and (2.) methodology. Scope refers to a study's relevance beyond the setting in which the research took place and denotes the applicability of conclusions to the broader population of which participants might reasonably form a representative sample: i.e. primary aged, Thai, students attending international schools. Scope depends primarily (although not exclusively) on the eligibility criteria employed to recruit from among hopeful study participants.

Those who did not participate in the current project fell into four categories: (1.) students from low socio-economic backgrounds, (2.) children on the EAL register, (3.) pupils receiving SEN support and, (4.) children deemed particularly able. Whether findings might apply to these groups therefore remains unclear. Children included on the SEN register did not participate because their responses to RR would plausibly have differed from those of their peers not receiving learning support (Section 4.7). The most able children did not take part for the same reason; to have included this latter group would have inflated estimations of learning outcomes beyond those likely from the 'average' child, potentially overestimating likely RR benefits for the more 'typical' student. The study did not involve children from low socio economic background simply because as a private, 'for profit,' institution the host school catered to the needs of affluent Thai and expatriate families. With these several categories of child removed, what remains of the school population are the average-performing Thai L1 speakers comprising the majority of the host school's roll. The dissertation's findings apply to this restricted, albeit numerically important, student grouping along with children in other Thai based international schools with a similar roll and student demographic profile.

Among the more contentious methodological concerns is whether researchers should resort to modified texts to explore vocabulary development. While Nagy et al. (1985, 1987) regarded non-authentic scripts as threats to ecological validity, others consider such concerns overemphasized or even misplaced (e.g., Swanborn & deGlopper, 1999). Complicating the issue is the question of what test of authenticity to apply. Most texts, after all, incorporate adaptations of some form to comply with curriculum specifications or perceived student needs. Textbook writers habitually simplify concepts with EAL children in mind or deliberately make use of the child's L1 in marginal glosses as a matter of course. As for school reading schemes, without adaptation, it could prove difficult compiling a collection of books that meet the requirement of i+1 (Section 2.4), while successfully engaging young readers. Not least, the term *adaptation* lies open to alternative interpretations, denoting almost anything from extensive rewriting of scripts to minimal amendments leaving a parent text little altered. Whether, and by how much, ecological validity diminishes given non-authentic texts most reasonably depends upon how clumsy, extensive and intrusive are the adaptations themselves. Just as adaptation is a matter of degree, so too, surely, any implications arising therefrom.

In any event, it arguably makes most sense to evaluate children's word gains from texts they most frequently engage with, whether authentic or otherwise. Given that unadapted reading materials feature only minimally among children's RR scripts, then findings based upon those same texts will likely prove rather less meaningful, at least from a teacher's perspective, than insights from research drawing upon adapted variants. Whether making use of adapted texts in the current research project amounted to a prudent researcher choice is a different issue altogether, the answer to which depends upon the practical difficulties that arise when exploring spaced learning in classroom settings, and the disruption to school routines that an authentic text based study might involve. For the purposes of the present study, experimental texts recommend themselves, as we have seen (Chapter 4), because they enable researchers to control for factors otherwise threatening the robustness of study conclusions. Indeed, it is difficult to see how any investigation employing unadapted reading materials would allow for isolating out target word presentation time from the effects of other potential determinants of word gains (e.g. helpfulness of contextual clues, word length, syntactic complexity to mention but a few). As for objections that experimental texts amount to unnatural or contrived artefacts, it is worth recalling that during the compilation process assessors (Section 4.6) evaluated each script for comparability with regular RR materials. Scripts judged unduly 'artificial' underwent revision to ensure they appeared more authentic to teachers and students alike.

A second methodological concern (Nagy et. al.,1985) relates to the probability of atypical reading behavior from student awareness that testing will follow on from their reading experiences. While an obvious concern –Nagy argues that readers will concentrate unduly on the text– the threat to 'design robustness' will, in practice, always depend upon the likelihood of atypical reading given the *particular* research setting, students, and study design. The question, viewed in this light, becomes how children regard, and respond, to testing as such. A child aware of an upcoming assessment may, or may not, read as s/he would if unaware of the post-reading assessment. Whether s/he modifies reading behavior depends upon his/her aversion to tests and importance s/he attaches to results –if, for example, s/he regards a particular test as threatening and anxiety raising as opposed to just another classroom-based activity. The issue, as regards the present study, resolves into how children regarded their VSAT sessions during their weeks of participation. The impression of myself and assistants was the same: that participants viewed the VSAT less as a test than a regular class task, not unlike others completed during the school day. VSAT sessions did not take place under conditions children associated with formal testing; absent were the cardboard partitions, the time limits, and teacher urgings that everyone "give of their best." On no occasion did all children complete the same VSAT sheet at the same time, as they would under typical school test conditions. Even after administrator marking, students seemed disinclined to ask how they had scored -none of the familiar queries regarding marks and possible grades.

A more statistical orientated concern arises with 'tied' scores and their treatment in binomial sign tests (Section 4.21.1). The study adopts a zeros 'workaround' which acknowledge zeros (i.e. tied pairs of data, identical in value) as evidence supportive of the null-hypothesis position. By allocating zeros in the manner described in Section 4.21.1, the study responds to Geyer's (2005) concern that researchers all too willingly engage in "honest cheating," i.e. designing methodologies intended to return significant p-values with little regard to how trivial the hypotheses the study aims to affirm. On the other hand, apportioning zeros as Geyer suggests, potentially conceals valuable findings given the greater likelihood that now arises of a Type 2 error. Interestingly, in several cases in which the study employed zero corrected tests, results could miss significance by the narrowest of margins (several examples appear in Part 2, Chapter 5). Had the current study not compensated for zeros, further instances of statistical significance become a real possibility.

The findings from any investigation prove only as robust as the data from which they derive. Like MCTs, and alternatives such as the VKS and State Rating Task, the VSAT comes with its own particular limitations. The original instrument proved time consuming to administer, prompting trials of a pen-and-paper alternative that dispensed with the researcher having to read out each target word (see Chapter 4); both test formats elicited (essentially) identical word-to-state assignments, the paper version, however, allowing for substantial time savings. The likelihood of children having ‘cheated’ while completing VSAT sheets (i.e. copied answers from a friend) seems ‘low’ given that a participant only rarely sat next to another engaged in completing the same test. At no point did the researcher or assistants observe a child copying the work of a ‘friend;’ nor at any time did a student report any such behavior among his or her peers. More generally, a case for VSAT use in school settings derives from the exhaustive testing of its parent, the VKS (see Paribakht & Wesche, 1993, 1996) along with results from three small-scale experiments (Chapter 3) that explored VSAT suitability for classroom research. With the benefit of hindsight, and having completed numerous VSAT tests during the current study, the researcher and assistants’ views remained as they had after participating in the VSAT validation studies (see Chapter 3): All agreed that the instrument seemed well-suited for eliciting reliable data from young, primary aged, EAL participants.

6.5 The need for further research

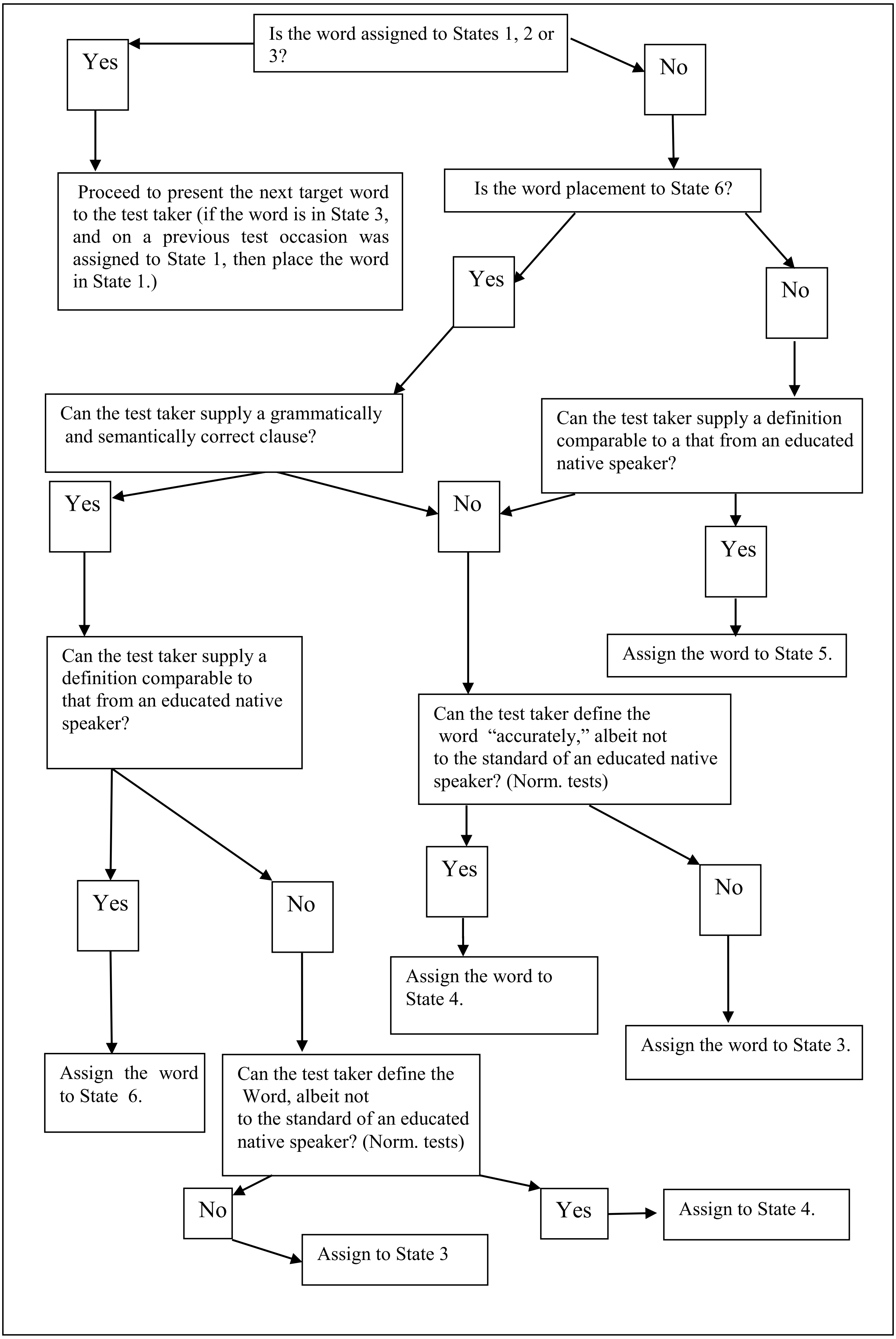
The dissertation establishes that the time intervals between encounters with the same novel word affect learning, yet barely touches upon what may prove a productive and interesting field of inquiry. Many intriguing questions remain: How might gains have differed had readers encountered more, or less, than 12 exposures to each target word? Would children have learned more from encountering target words over 6 days, 7 or 8? How does the interval over which learning occurred affect the durability of gains over weeks, months or even years? And, not least, how might results have differed had the research involved secondary or tertiary level students as opposed to those in primary? It would prove interesting, and feasible, to employ the same methodology with children from non-Thai L1 backgrounds or with English native speakers in an effort to identify general findings applicable irrespective of students’ L1, age, and/or cultural background. Extensions to the current work might also reasonably involve replicative studies involving participants with needs that set them aside from the typical student population. Children with dyslexia, attention deficit disorder, and elective mutism, for example, may learn and acquire language atypically from the norm but nevertheless gain vocabulary from reading experiences given appropriate texts and a conducive reading environment. To replicate the present study with such children should not prove unduly difficult after refining VSAT administration procedures as seem prudent –perhaps with input from the Special Educational Needs Department. Nor does it seem impractical to investigate spaced learning effects upon readers singled out as particularly able, whether EAL students or English native speakers. Perhaps most importantly, however, lies the need for corroborative research. Ideally, this might employ rather different methodologies to those the current study adopted, perhaps making use of unadapted reading texts and students from other than Year 4 classes. Such validation would supply a powerful rationale and justification to move beyond the issues addressed in the present investigation and explore other contributions of spaced learning to lexical gains from RR experiences.

Although the dissertation examines vocabulary expansion from reading, the methodology lends itself to exploring spaced learning effects from listening experiences (for example, from children hearing stories read by the class teacher) or even concomitant reading and listening sessions of the sort Horst (2001) introduced to her tertiary level EAL learners. From a practicing teacher's perspective it makes sense to broaden the ambit of spaced learning research to determine the impact of intervals between word reencounters from non-recreational reading experiences such as children engage in to learn topic content related to, say, history, science and mathematics. A further line of inquiry might involve studies that set out to identify the effects of spaced target word encounters on aspects of vocabulary knowledge other than word–meaning associations (e.g., spelling proficiency).

6.6 Conclusion

In the several decades since Saragi, Nation, and Meister (1978) demonstrated that readers gain word meanings from recreational reading, the body of literature on incidental vocabulary research has expanded into a substantial collection of articles, books, and technical reports, all claiming to shed light on factors impacting upon learning outcomes. Although this research has yielded pedagogically significant findings, at present no comprehensive model exists that satisfactorily explains the depth and breadth of word gains from recreationally engaging with texts. The current study provides research-backed evidence that a previously unexplored factor, the time interval between a reader's encounters with novel words, contributes to the depth and breadth of vocabulary expansion from RR sessions. The study demonstrates that these intervals assist in our understanding of incidental word gains and that credible models of how such gains arise must acknowledge the 'spacing' factor among other relevant variables. Whether textual manipulation to exploit the spacing effect proves practical, or even desirable, depends upon the significance we attach to anticipated learning gains weighed against the non-trivial challenges that textual adaptation entails. It remains the researcher's hope that this investigation encourages further studies into this intriguing issue –that linguists and teachers will explore the costs and benefits of designing reading materials with a view to optimizing texts for children's vocabulary expansion and general literacy development.

Appendix 1 Guide to administering the VSAT



Appendix 2a Sums of words children knew to the standard of VSAT state 6

Table 1 (State 6)

	Set 1	Set 2	Set 3	Set 4	Set 5	Total
1	n	n	n, adj	n	n	6
2	n	v, adj	n, adj		n,v	7
3	n, adv	n, adv	n, adj,adv	n	n	9
4	v	v	v	v	adj	5
5	adj	n, adj		n,v, adj	n, adv	8
6	adj	n	n		n	4
7	n	n		v, adj	adj	5
8	adj	adj	n	n	n,v	6
9	n, adj	adj			n	4
10	n	n	n	n	n	5
11		v	adj		n, adv	4
12	n	adj			v	3
13		n, adv	adj, adv	n, adv	n, adv	8
14	n,v				n	3
15		v	adj	v		3
16	v, adj	n	n, adj		adj	6
17	n	n	v	adj, adv	n	6
18	n	n, adv	n, adj, adv	adj		7
19				n, adv	n, adv	4
20	n			n		2
21	adj	n	adj			3
22		n,v	n	adj	v	5
23	n,v			n, adj	v	5
24	n			n	n, adj	4
25			n	v	adj	3
26	adv			adj	n	3
27				n	adv	2
28	n,adv			n	v	4
	27	25	24	27	31	

Appendix 2b Sums of words children knew to the standard of VSAT State 6+5

Table 2 (State 6+5)

	Set 1	Set 2	Set 3	Set 4	Set 5	Total
1	n,adj	n	n,v,adj	n,adj	n,v,adj	11
2	n,v	n,v,adj	n,adj,adv	n	n,v,adj,adv	13
3	n,adv,adj	n,adv	n,v,adj,adv	n,v,adv	n,adv	14
4	n,v	v,adj	v,adj,adv	v,adj	n,adj,adv	12
5	n,adj,adv	n,v,adj	n,v,adv	n,v,adj,adv	n, v,adj,adv	17
6	adj	n,adj	n	n,v	n, v, adj	9
7	n	n		n,v,adj,adv	n, adj	8
8	n, adj, adv	n, adj, adv	n,v	n,v	n, v, adj	13
9	n, adj	adj, adv	v, adv	adj	n, v, adj	10
10	n	n, adv	n, v	n, v	n, v	9
11	v, adj	v	n, adj		n, adj, adv	8
12	n, v, adj	adj, adv	n, v	n, adv	n, v, adv	12
13	n	n, adj, adv	adj, adv	n, v, adv	n, adv	11
14	n, v, adj		n	v, adj	n, v, adj	9
15		v, adj	n, adj	n, v, adj	n	8
16	v, adj	n, v	n, adj	n, adj	n, adj	10
17	n	n	n, v, adv	v, adj, adv	n, v	10
18	n,v	n, adv	n,v,adj, adv	n, adj	n	11
19	n	adj	n,v,adv	n, adv	n,v,adj,adv	11
20	n	adj	adv	n, v, adj	n, adj	8
21	n, adj	n,v	adj	adj, adv	n, v	9
22	n, adv	n, v	n, v	n,v,adj	n,v	11
23	n, v, adj	n, adv	v	n, v, adj	n, v	11
24	n,	v, adj	n,adj	n,v,adj	n, adj, adv	11
25	n, v	n, v	n, v	n, v, adv	n, v, adj	12
26	adv			n, adj	n, adj	5
27	adj	adj		n,adv	v, adj, adv	7
28	n, v, adv	n, v	n, v	n, adj	n, v	11
	51	49	55	65	71	

Appendix 2c Sums of words children knew to the standard of VSAT States 6+5+4

Table 3 (State 6+5+4)

	Set 1	Set 2	Set 3	Set 4	Set 5	Total
1	n,v,adj	n,adj,adv	n, v, adj	n,v,adj,adv	n,v,adj	16
2	n,v,adj	n,v,adj	n, v,adj,adv	n,v	n,v,adj,adv	16
3	n, v,adv,adj,	n,adv	n, v,adj,adv	n,v,adv	n,v,adv	16
4	n,v	v,adj	n,v,adj,adv	v,adj	n,v,adj,adv	14
5	n,adj,adv	n,v,adj	n,v,adj,adv	n,v, adj,adv	n, v,adj,adv	18
6	v,adj	n,adj	n,v,adj	n,v,adv	n,v,adj,adv	14
7	n,adv	n,adj	n,adj	n,v, adj,adv	n,v,adj	13
8	n,adj,adv	n,adj,adv	n,v,adv	n,v	n,v,adj	14
9	n, v,adj	v,adj,adv	n,v,adv	n,adj,adv	n,v,adj	15
10	n	n,v,adj,adv	n,v,adv	n,v,adj	n,v	13
11	n,v,adj	n,v,adj	n,v,adj	n	n, v,adj,adv	14
12	n,v,adj,adv	v,adj,adv	n,v,adj	n,adj,adv	n,v,adj,adv	17
13	n	n,adj, adv	v,adj, adv	n, v,adv	n, adv	12
14	n,v,adj,adv	n,adj	n,adv	v,adj	n,v,adj	13
15	adj	v,adj	n,adj	n,v,adj	n,v,adv	11
16	v, adj	n,v,adj	n, v,adj	n,v,adj	n,v,adj,adv	15
17	n,adv	n,v	n,v,adv	v,adj, adv	n,v,adj	13
18	n,v	n, v,adv	n, v,adj, adv	n,v,adj	n,v,adj	15
19	n,adj	v,adj	n,v,adj,adv	n, adv	n, v,adj,adv	14
20	n,v,adv	adj	n,v,adv	n,v,adj	n,adj,adv	13
21	n, adj,adv	n,v	n,adj,adv	v,adj,adv	n,v,adj	14
22	n,adv	n,v	n,v,adj,adv	n,v,adj	n,v,adv	14
23	n,v,adj,adv	n,v,adv	v,adj	n, v,adj,adv	n,v,adj	16
24	n,	v,adj,adv	n,adj	n,v,adj	n,v,adj,adv	13
25	n,v	n,v,adj	n,v,adj	n,v,adv	n,v,adj,adv	15
26	v,adv	n	v,adv	n,v,adj	n,adj,adv	11
27	v, adj	n,v,adj		n,v,adv	n,v,adj,adv	12
28	n,v,adv	n,v,adj,adv	n,v	n,adj	n,v	13
	69	72	81	80	92	

Appendix 3a VSAT test for set 1

Name__________________ Date__________________

Match the words to the sentences (1)

I don't know this word!			*trop* *learn* *try* *dip* *lazily* *problem* *ape* *cup* *nish* *rubber* *harg* *write* *happily* *photo* *rendly* *bath* *uncle* *cruel* *laugh* *easily*
I haven't seen this word but I think I know what it means!			
I have seen this word before but I don't know what it means!			
I have seen this word before and I think it means…			
I know this word, It means …			
I can use this word in a sentence.			

Time started ………………….. Time finished ……………………….

<u>Appendix 3b</u> VSAT test for set 2

Name__________________ Date__________________

Match the words to the sentences (2)

I don’t know this word!			*torg* *cake* *kind* *ball* *doctor* *brush* *heavy* *ned* *clip* *blood* *yellow* *trag* *mouse* *keypad* *torly* *smelly* *sadly* *knife* *catch* *cheekily*
I haven’t seen this word but I think I know what it means!			
I have seen this word before but I don’t know what it means!			
I have seen this word before and I think it means…			
I know this word, It means …			
I can use this word in a sentence.			

Time started ………………….. Time finished ……………………….

<u>Appendix 3c</u> VSAT test for set 3

Name__________________ Date__________________

Match the words to the sentences (3)

<table>
<tr><td>I don't know this word!</td><td></td><td></td><td rowspan="12">snatch
play
growl
powl
slime
tep
jumper
gos
breakfast
gain
altitude
trip
gleefully
pill
turn
sneply
frog
bowl
chap
pull</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I haven't seen this word but I think I know what it means!</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I have seen this word before but I don't know what it means!</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I have seen this word before and I think it means…</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I know this word, It means …</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I can use this word in a sentence.</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

Time started ………………….. Time finished ………………………

<u>Appendix 3d</u>   VSAT test for set 4

Name__________________   Date__________________

Match the words to the sentences (4)

I don't know this word!			*wobble* *pril* *liver* *snarl* *button* *slither* *sol* *shield* *crint* *shallow* *jellyfish* *trapeze* *happily* *clay* *garply* *frame* *creepily* *cruel* *squeeze* *caringly*
I haven't seen this word but I think I know what it means!			
I have seen this word before but I don't know what it means!			
I have seen this word before and I think it means…			
I know this word, It means …			
I can use this word in a sentence.			

Time started ………………….. Time finished ………………………

<u>Appendix 3e</u> VSAT test for set 5

<table>
<tr><td colspan="4">Name__________________ Date__________________

Match the words to the sentences</td></tr>
<tr><td>I don't know this word!</td><td></td><td></td><td rowspan="12">wost
pyramid
tidy
easy
drink
lont
cloud
larb
desk
count
square
tick
vision
parn
nail
smart
shirt
cleanly
ride
chew</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I haven't seen this word but I think I know what it means!</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I have seen this word before but I don't know what it means!</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I have seen this word before and I think it means…</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I know this word, It means …</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td>I can use this word in a sentence.</td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td colspan="4">Time started ………………….. Time finished ………………………</td></tr>
</table>

Appendix 4 Sums of words occupying VSAT State 1

Case Summaries

	Set.1	Set.2	Set.3	Set.4	Set.5
1	1.00	1.00	.00	.00	1.00
2	1.00	.00	.00	.00	.00
3	.00	2.00	.00	.00	.00
4	2.00	1.00	.00	.00	.00
5	.00	1.00	.00	.00	.00
6	1.00	.00	1.00	1.00	.00
7	1.00	.00	.00	.00	1.00
8	.00	.00	.00	.00	.00
9	1.00	.00	1.00	.00	.00
10	1.00	.00	1.00	.00	1.00
11	1.00	.00	.00	1.00	.00
12	.00	1.00	1.00	.00	.00
13	2.00	.00	.00	1.00	1.00
14	.00	2.00	.00	.00	.00
15	1.00	1.00	2.00	.00	1.00
16	1.00	.00	.00	.00	.00
17	1.00	1.00	1.00	1.00	.00
18	1.00	.00	.00	1.00	.00
19	.00	.00	.00	1.00	.00
20	1.00	1.00	.00	.00	.00
21	.00	1.00	1.00	.00	1.00
22	1.00	2.00	.00	.00	.00
23	.00	1.00	.00	.00	.00
24	1.00	1.00	1.00	.00	.00
25	1.00	.00	.00	.00	.00
26	.00	1.00	1.00	.00	1.00
27	2.00	.00	1.00	1.00	.00
28	.00	.00	1.00	1.00	.00
Sum	21.00	17.00	12.00	8.00	7.00

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Ames, W. S. (1966). The development of a classification scheme of contextual aids. *Reading Research Quarterly, 2*, 57-82.

Anders, P., & Bos, C. (1986). Semantic feature analysis: An interactive strategy for vocabulary development and text comprehension. *Journal of Reading*, *29*, 610-616.

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

Anderson-Clarke, T. N., Green, R. J., & Henley, T. B. (2008). The relationship between first names and teacher expectations for achievement motivation. *Journal of Language and Social Psychology*, *27*(1), 94-99.

Aston, G., & Burnard, L. (1998). The BNC handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Atkinson, A. (1987). The mother tongue in the classroom: A neglected resource? *ELT Journal*, *41*(4), 241-247.

Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(2), 126-133.

Atkinson, R. C., & Shifrin, R.M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (pp. 89-195). New York: Academic Press.

Au, T. K., & Glusman, M. (1990). The principle of mutual exclusivity in word learning: To honor or not to honor? *Child Development*, *61*(5), 1474-1490.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology*: *General*, *113,* 1-2.

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 344-349.

Banaji, M.R., & Crowder, R.G. (1989). The bankruptcy of everyday memory. *American Psychologist. 44*, 1185-1193.

Barnett, M. (1989). Teaching reading strategies: How methodology affects course articulation. *Foreign Language Annals*. *21*, 109-21.

Baron-Cohen, L. A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, *21*, 37-46.

Bartlett, F. (1932). *Remembering: An experimental and social study*. Cambridge, UK: Cambridge University Press.

Baumann, J. F., Kameenui, E. J, & Ash, G. (2003). Research on vocabulary instruction: Voltaire redux. In D. Lapp, J. R. Squire, & J. Jensen (Eds.), *Handbook of research on teaching the English language arts* (pp. 752-785). Mahwah, NJ: Lawrence Erlbaum.

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, *83*(3), 177-181.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). The effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74*, 506-521.

Becker, W. C. (1977). Teaching reading and language to the disadvantaged: What we have learned from field research. *Harvard Educational Review*, *47*, 518-543.

Behrend, D.A. (1990). Constraints and development: A reply to Nelson (1988). *Cognitive Development*, *5*, 313-330.

Benjamin, A., & Crow, J. T. (2010). *Vocabulary at the center*. Larchmont, NY: Eye On Education.

Biedalak, K., & Wozniak, P. A. (1996). Build your vocabulary with SuperMemo. *Spotlight*, *10,* 4.

Birdsong, D. (1999). Introduction: Whys and why nots of the critical period hypothesis. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 1-22). Mahwah, NJ: Erlbaum.

Blachowicz, C. L. Z., & Fisher, P. (2006). *Teaching vocabulary in all classrooms* (2nd ed.). Columbus, OH: Prentice-Hall.

Bley-Vroman, R. (2009). The evolving context of the fundamental difference hypothesis. *Studies in Second Language Acquisition*, *31*, 175-198.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bloom, P. (2001). Word learning. *Current Biology*, *11*, 5-6.

Bloom, P. (2004). Myths of word learning. In D. G. Hall & S.R. Waxman (Eds.), *Weaving a Lexicon*, (pp. 205-224). Cambridge, MA: MIT press.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, *74*, 245–248.

Broadbent, D. (1958). Perception and communication. Oxford: Oxford University Press.

Bruton, A. (2009). The Vocabulary Knowledge Scale: A critical analysis. *Language Assessment Quarterly*, *6*(4), 288-297.

Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. *Child Development*, *61*, 201-210.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1-47.

Carmines, E.G., & Zeller, R.A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369-378.

Carr, T. H., & Curran, T. (1994). Cognitive factor s in learning about structured sequences: Applications to syntax. *Studies in Second Language Acquisition*, *16*,205-230.

Carrell, P. L. (1983). Three components of background knowledge in reading comprehension. *Language Learning*, *33*, 183-207.

Carroll, S. E. (2006). Salience, awareness and SLA. In M. G. O'Brien, C. Shea, & J. Archibald (Eds.), *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference* (GASLA 2006) (pp. 17-24). Somerville, MA: Cascadilla Proceedings Project.

Carver, R. P. (1983). Is reading rate constant or flexible? *Reading Research Quarterly*, *18*(2), 190-215.

Carver, R. P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior*, *26*(4), 413-437.

Challis, B. H. (1993). Spacing effect on cued-memory tests depend upon level of processing. *Journal of Experimental Psychology, Learning, Memory, and Cognition*, *19*, 389-396.

Cheng, H. C. (2011). *Vocabulary acquisition in learning English as a second language: Examining the involvement load hypothesis and language anxiety with Taiwanese college students* (Doctoral dissertation, University of Northern Colorado).

Chiswick, B., & Miller, P. (2004). *Linguistic distance: A quantitative measure of the distance between English and other languages*. Discussion Paper Series. Institute for the Study of Labor.

Cho, K. S., & Kim, H.-J. (2004). Recreational reading in English as a foreign language in Korea: Positive effects of a sixteen-week program. *Knowledge Quest, 32*(4).

Cho, K. S., Park, E. J., & Krashen, S. (2008). Reading in English by children in Korea: Frequency, effectiveness and barriers. *Indonesian JELT, 4*(1), 18-25.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *3*, 113-124.

Chomsky, N. (1957). *Syntactic structures.* Mouton and Co., The Hague.

Chomsky, N. (1959). Review of Skinner's Verbal Behavior. *Language*, 35, 26–58.

Chua, S. (2008). The effects of the sustained silent reading program on cultivating students' habits and attitudes in reading books for leisure. *Clearing House, 81*(4), 180.

Clark, E. V. (1983). Meanings and concepts. In J. H. Flavell, E. M. Markman, & P. H. Mussen (Eds.), *Handbook of child psychology*: *Vol. 3. Cognitive development* (pp. 787-840). New York: Wiley.

Cohen, M. S., Yan, V. X., Halamish, V., & Bjork, R. A. (2013, April 8). Do Students Think That Difficult or Valuable Materials Should Be Restudied Sooner Rather Than Later? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. doi:10.1037/a0032425.

Coles, M. (1982). *Word perception, first language script and learners of English as a second language* (M.A. dissertation, University of London, Birkbeck College, London).

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204-256.

Coxhead, A. (2010). Grabbed early by vocabulary: Nation's ongoing contributions to vocabulary and reading in a foreign language. *Reading in a Foreign Language*, *22*, 1-14.

Craik, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 675-676.

Craik, F., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology*, *104*, 268-294.

Craik, F. I. M. (2002). Levels of processing: Past, present and future? *Memory*, *10*, 305-318.

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research*, *36*, 206-217.

Crowell, S. E. (2004). The neurobiology of declarative memory. In J. H. Schumann, S.E. Crowell, N.E. Jones, & S. A. Schuchert (Eds.), *The neurobiology of learning: Perspectives from second language acquisition*. London: Lawrence Erlbaum Associates.

Cunningham, A. E. (2005). Vocabulary growth through independent reading and reading aloud to children. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice.* Mahwah, NJ: Erlbaum.

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

Dempster, F. N. (1988). Informing classroom practice: What we know about several task characteristics and their effects on learning. *Contemporary Educational Psychology*, *13*(3), 254-264.

Denham, K., & Lobeck, A. (2009). *Linguistics for everyone.* Boston, MA: Wadsworth Publishing.

DeWitt, J. E. (1994). How a word-learning principle might interact with metalinguistic knowledge and input. In *The Proceedings of the Annual Child Language Research Forum* (Vol. 26, p. 159). Stanford Linguistics Association.

Diakidoy, I. N. (1998). The role of reading comprehension in word meaning acquisition during reading. *European Journal of Psychology of Education*, *13*, 131-154.

Dollaghan, C. A. (1985). Child meets word: Fast mapping in preschool children. *Journal of Speech and Hearing Research*, *28*, 449-454.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect. *Journal of Applied Psychology*, *84*, 795-805.

Dore, J. (1978). Conditions for the acquisition of speech acts. In L. Markova (Ed.), *The social context of language* (pp. 87-111). New York, NY: Wiley.

Dresser, C. (1994). Shade for sale: A Chinese tale. In *The rainmaker's dog* (pp. 223-234). New York: St. Martin's Press.

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, *120*, 190-202.

Ebbinghaus, H. (1895/1964). *Memory: A contribution to experimental psychology*. Mineola, NY: Dover Publications.

Elhoweris, H., Mutua, K., Alsheikh, N., & Holloway, P. (2005). Effect of children's ethnicity on teacher's referral and recommendation decisions in gifted and talented programs. *Remedial and Special Education*, *26*(1), 25-31.

Elley, W. B. (1988). New vocabulary: How do children learn new words? *Research Information for Teachers, 1*, 1-4.

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly, 24,* 174-186.

Elley, W. B. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language Learning, 41*, 375-411.

Ellis, N. C. (1994). Consciousness in second language learning: Psychological perspectives on the role of conscious processes in vocabulary acquisition. *AILA Review*, *11*, 37-56.

Ellis, N. C. (2000) Memory for Language. In P. Robinson (Ed.) *Cognition and Second Language Instruction,* pp. 33-68. Cambridge: Cambridge University Press.

Ellis, N. C. (1995). The psychology of foreign language acquisition: Implications for CALL. *International Journal of Computer Assisted Language Learning* (CALL), *8*, 103-128.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition 27*, 305-352.

Ellis, N. C. (2008). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal*, *92*, 232-249.

Eysenck, M., & Keane, M. T. (2005). *Cognitive psychology: A student's handbook.* Philadelphia, PA: Taylor and Francis.

Faruji, L. F. (2012). Declarative vs procedural memory: Roles in second language acquisition. *Broad Research in Artificial Intelligence and Neuroscience, 3*(1), 37-49.

Fielding, L. G., Wilson, P.T., & Anderson, R. C. (1989). A new focus on free reading: The role of trade books in reading instruction. In T. E. Raphael & R. Reynolds (Eds.), *Contexts of literacy.* White Plains, NY: Longman.

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling. *Journal of Educational Psychology, 59*, 290-296.

Flavell, J. (1988). The development of children's knowledge about the mind: From cognitive connections to mental representations. In J. Astington, P. Harris, & D. Olson (Eds.), *Developing theories of mind* (pp. 244-267). Cambridge: Cambridge University Press.

Ganger, J., & Brent, M. (2001). Re-examining the vocabulary spurt and its implications. Is there really a sudden change in cognitive development? In A. H. Do, L. Dominguez, & A. Johansen (Eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Ganger, J., & Brent, M. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*, 621-632.

Gao, H. (2013). A study on constructing sense relations. *Sino-US English Teaching*, *10*(5), 366-372.

Garan, E., & DeVoogd, G. (2008). The benefits of sustained silent reading: Scientific research and common sense coverage. *The Reading Teacher*, *62*(4).

Garan, A. M. (2008) Cultural competence, racial identity, and ethnocultural empathy in therapeutic engagement with at risk, urban adolescents of color. (Doctoral dissertation, Smith College School for Social Work). Dissertation Abstracts International.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied Linguistics, 25*, 1-37.

Gass, S. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, *9*, 198-217.

Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S. M., & Selinker, L. (2001). *Second language acquisition: An introductory* *course*. London: Lawrence Erlbaum.

Genesee, F., & Nicoladis, E. (2006). Bilingual first language acquisition. In E. Hoff & M. Shatz (Eds.), *Handbook of language development* (pp. 324-342). Oxford, England: Blackwell Publishing.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development* (pp. 301-334). Hillsdale, NJ: Erlbaum.

Geyer. (2013). Statistics, 5601, Sign test. Retrieved from: http://www.stat.umn.edu/geyer/5601/examp/sign.html

Gibbons, P. (2002). *Scaffolding Language, Scaffolding Learning*. Portsmouth NH: Heinemann.

Gleitman, L. (1990). Structural sources of verb learning. *Language Acquisition*, *1*, 1-63.

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

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11*, 341-363.

Griffiths, C. (2008). Strategies and good language learners. In C. Griffiths (Ed.), Lessons from good language learners (pp. 83-98). Cambridge: Cambridge University Press.

Hafiz, F., & Tudor, I. (1990). Graded readers as an input medium in L2 learning. *System*, *18*(1), 31-42.

Hakuta, K. (1974). A preliminary report on the development of grammatical morphemes in a Japanese girl learning English as a second language. *Working Papers on Bilingualism*, *3*, 18-38.

Harley, B. (1992). Patterns of second language development in French immersion. *Journal of French Language Studies*, *2*, 159-183.

Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition*, *19*, 379-400.

Harley, B., & Wang, W. (1997). The critical period hypothesis: Where are we now? In A. M. B. de Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives* (pp. 26-31). Hillsdale, NJ: Lawrence Erlbaum.

Harris, R. J. (2001). *A primer of multivariate statistics*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experiences* *of young American children*. Baltimore, MD: Brookes Publishing.

Hartshorne, J. K., & Ullman, M. T. (2006). Why girls say “holded” more than boys. *Developmental Science, 9,* 21-32.

Hasher, M., & Chromiak, W. (1977). The processing of frequency information: An automatic mechanism? *Journal of Verbal Learning and Verbal Behavior, 16*, 173-184.

Heimlich, J. E., & Pittelman, S. D. (1986). Semantic mapping: Classroom applications. Reading Aids Series, IRA Service Bulletin. *International Reading Association*. Newark, NJ: Heimlich & Pittelman.

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

Hill, D., & Thomas, H. R. (1988). Survey of graded readers. *ELT Journal*, *42*(1), 44-52.

Hintzman, D.L. (1976). Theoretical implications of the spacing effect. In R.L. Solso (Ed.), *Theories in cognitive psychology*: *The loyola symposium.* Potaomac, Maryland: Lawrence Earlbaum Associates.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8,* 689-696.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods*. New York: John Wiley & Sons.

Horst, M. (2001). *Text encounters of the frequent kind: Learning L2 vocabulary through reading.* (Unpublished doctoral dissertation, University of Wales, Swansea, UK).

Horst, M., Cobb T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207-223.

Howatt, A. P. R. (1984). *A history of English language teaching*. Oxford: Oxford University Press.

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403-430.

Huang, Y. (2004). Testing the involvement load hypothesis: Effects of reading-based tasks on vocabulary retention by Chinese EFL learners. *Journal of Modern Foreign Languages*, *4*, 4.

Huitt, W. (2003). The information processing approach to cognition.

Educational Psychology Interactive. Valdosta, GA: Valdosta State University.

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: Current perspectives on teaching the four skills. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258-287). Cambridge: Cambridge University Press.

Hulstijn, J. H. (1989). Implicit and incidental second language learning: Experiments in the processing of natural and partly artificial input. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processing*, (pp. 49-73). Tubingen: Gunter Narr.

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), 327-339.

Huttenlocher, J. (1974). The origins of language comprehension. In R. Solso (Ed.), *Theories in cognitive psychology* (pp. 331-368). New York, NY: Erlbaum.

Ijaz, I. H. (1986). Linguistic and cognitive determinants of lexical acquisition in a second language. *Language Learning*, *36*(4), 401-451.

Jackendoff, R. (1987). *Consciousness and the computational mind*. New York, NY: Academic Press.

Jackson, G. M., & Jackson, S. R. (1995). Do measures of explicit learning actually measure what is being learned in the serial reaction time task? A critique of current methods. *Psyche, 2*(20).

Jacobs, M. G., Dufon, P., & Hong, C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, 17,19-28.

James, W. (1890). The Principles of Psychology. In F. H. Burkhardt, F.Bowers & I. K. Skrupskelis (Eds). *The Works of William James*. Cambridge: Harvard University Press.

Jamieson, P. (1976). *The acquisition of English as a second language by young Tokelau children living in New Zealand* (Unpublished Ph.D. thesis, Victoria University of Wellington, Wellington, New Zealand).

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, *30*, 138-149.

Janopoulous, M. (1986). The relationship of pleasure reading and second language writing proficiency. *TESOL Quarterly*, *20*, 764-768.

Jedynak, M. (2009). *The critical period hypothesis revisited: The impact of age on ultimate attainment in the pronunciation of a foreign language.* Frankfurt: Lang.

Jenpattarakul, W. (2012). The impact of keyword technique on the students' vocabulary retention ability in an EFL class. *Mediterranean Journal of Social Sciences, 3*(3), 265-57.

Joe, A. (1995). Text-based tasks and incidental vocabulary learning. *Second Language Research*, *11*, 149-158.

Joe, A. (1998). What effects do text-based tasks promoting generation have on vocabulary acquisition? *Applied Linguistics*, *19*, 357-377.

Jourdenais, R. (1998). *The effects of textual enhancement on the acquisition of the Spanish preterit and imperfect* (Unpublished Ph.D. dissertation, Georgetown University, Washington, DC).

Kachroo, J. N. (1962). Report on an investigation into the teaching of vocabulary in the first year of English. *Bulletin of the Central Institute of English*, *2*, 67-72.

Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple choice tests. *Journal of Memory and Language*, *57*, 151.

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.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning*, *58*(2), 285-325.

Kirby, M. C. (2003). The effects of weekly, sustained silent reading time on recreational reading habits and attitudes in a 9th grade English class. Retrieved from ERIC database. (ED479485).

Klump, J. (2007). Knowledge and input: Bilingual preschoolers' use of the mutual exclusivity assumption. Papers and reports on child language development. *Proceedings of the Child Language Research Forum*, *26*, 159-166.

Kolers, P. A., & Ostry, D. J. (1974). Time course of loss of information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior*, *13*(6), 599-612.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the enemy of induction? *Psychological Science*, *19*(6), 585-592.

Kornell, N., Bjork, R. A., Castel, A.D. & Eich, T.S. (2010). Spacing as the friend of both memory and indction in young and older learners. *Psychology and Aging* (2010), *25*, 498-503.

Krashen, S. (1987). *Principles and practice of second language acquisition*. New York: Prentice Hall.

Krashen, S. (2004). *The power of reading: Insights from the research*. Portsmouth, NH: Heinemann.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Language experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606-608.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychologist*, *6*, 293-323.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds.), *Vocabulary and Applied Linguistics* (pp.126-132). Basingstoke & London: Macmillan.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20-34). United Kingdom: Cambridge University Press.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694-716.

Laufer, B., & Hulstijn, J. H. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*, 3.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.

Laufer, B., & Sim, D. D. (1985). Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals*, *18*(5), 405-411.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York, NY: John Wiley & Sons.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, *26*(1), 35-57.

Lindsey, R.V., Shroyer, J.D., Pashler, H, & Mozer, M.C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*, 639-647.

Liu, C. K. (2007). What role self-selected reading can play in freshman English classes. *The Proceedings of 2007 International Conference and Workshop on TEFL & Applied Linguistics*, pp. 663-666.

Liu, J. (2009). L1 use in L2 vocabulary learning: Facilitator or barrier? *International Education Studies*, *1*(2), 65-69.

Lockhart, R. S., Craik, F. I. M., & Jacoby, L. L. (1976). Depth of processing, recognition and recall: Some aspects of a general memory system. In J. Brown (Ed.), *Recall and recognition*. London: Wiley.

Lockhart, R. S., & Craik, F. I. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology*, *44*, 87-112.

Lum, J., & Kidd, E. (2012). An examination into the associations between multiple memory systems, past tense and vocabulary in typically developing 5-year-old children. *Journal of Speech, Language and Hearing Research*, *55*, 989-1006.

Lyster, R. (1990). The role of analytic teaching in French immersion programs. The *Canadian Modern Language Review*, *47*, 159-176.

Ma, T. (2008). *The effectiveness of guessing vocabulary meaning using contextual clues and learning word lists in terms of vocabulary retention* (Unpublished Master's dissertation, University of Hong Kong, Hong Kong).

Mack, M. (2003). The phonetic systems of bilinguals. In M. T. Banich & M. Mack (Eds.), *Mind, brain, and language: Multidisciplinary perspectives* (pp. 309-349). Mahwah, NJ: Lawrence Erlbaum Press.

Marinova-Todd, S. (2003). *Comprehensive analysis of ultimate attainment in adult second language acquisition* (Unpublished doctoral dissertation, Harvard University, Harvard).

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199-227.

Martínez-Fernández, A. (2008). Revisiting the involvement load hypothesis: Awareness, type of task and type of item. In M. Bowles., R. Foote, S. Perpiñán, & R. Bhatt (Eds.), *Selected Proceedings of the 2007 Second Language Research Forum* (pp. 210-228). Somerville, MA: Cascadilla Proceedings Project.

Mason, B. (2007). The efficiency of self-selected reading and hearing stories on adult second language acquisition. *Selected Papers from the Sixteenth International Symposium on English Teaching, English Teachers' Association*. Republic of China, Taipei: Crane Publishing Company.

Mattioli, G. (2004). On native language intrusions and making do with words: Linguistically homogeneous classrooms and native language use. *English Teaching Forum*, *42*(4), 20-25.

McCarthy, D. (1954). Language development in children. In L. Carmichael (Ed.), *Manual of child development* (pp. 492-630). New York: Wiley.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. Ney York, NY: Appleton-Century-Crofts.

McCracken, R.A. (1971). Initiating Sustained Silent Reading. *Journal of Reading*, *14*, 521-524.

McShane, J. (1979). The development of naming. *Linguistics*, *17*, 879-905.

Meara, P. M., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142-154.

Merritt, S. (2006). *Mastering multiple choice: The definitive guide for achieving breakthrough results on multiple choice exams*. Creemore, ON: The Brain Ranch.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. New York: Multilingual Matters.

Möhring, A. (2001). The acquisition of French by German pre-school children. An empirical investigation of gender assignment and gender agreement. In S. Foster-Cohen & A. Nizegorodcew (Eds.), EUROSLA Yearbook 1, 171-193. Amsterdam: John Benjamins.

Mondria, J. A., & Mondria-De Vries, S. (1994). Efficiently memorizing words with the help of vocabulary cards and "hand computer": Theory and applications. *System*, *22*, 47-57.

Moonen, M., De Graaff, R., Westhoff, G., & Admiraal, W. (2005). Task effects on the acquisition of L2 vocabulary. Paper presented at the *2005 International Conference on Task-Based Language Teaching*. Leuven, Belgium.

Moore, R. K., & Louis ten Bosch, L. (2009). Modelling vocabulary growth from birth to young adulthood. In *INTERSPEECH* (pp. 1727-1730).

Morgan, G., & Kegl, J. (2006). Nicaraguan sign language and theory of mind: The issue of critical periods and abilities. *Journal of Child Psychology and Psychiatry*, *47*, 811-819.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal, 24*(2), 237-70.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*(2), 233-53.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.

Nation, I. S. P., & Heatley, A. (1994). Range: A program for the analysis of vocabulary in texts [software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.

Nation, I. S. P. (2000). Vocabulary. In M. Byram (Ed.) *Routledge encyclopaedia of language teaching and learning* (pp. 665-667). London: Routledge.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P., & Coady, J. (1988). Vocabulary and reading. In R. Carter & M. J. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 97-110). New York, NY: Longman.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*, 12-25.

National Reading Panel (US), National Institute of Child Health & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development, National Institutes of Health.

Nemati, A. (2009). Memory vocabulary learning strategies and long-term retention. *International Journal of Vocational and Technical Education*, *1*(2), 14-24.

Newton, M. (2002). *Savage girls and wild boys: A history of feral children*. London: Faber and Faber.

Nisbet, S. D. (1941). The scientific investigation of spelling instruction in Scottish schools. *British Journal of Educational Psychology*, *9*, 150.

Okanda, M., & Itakura, S. (2007). Do Japanese children say "yes" to their mothers? A naturalistic study of response bias in parent–toddler conversations. *First Language*, *27*, 421-429.

Okanda, M., & Itakura, S. (2008). Children in Asian cultures say yes to yes–no questions: Common and cultural differences between Vietnamese and Japanese children. *International Journal of Behavioral Development*, *32*, 131-136.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. New York, NY: Cambridge University Press.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, *45*, 255-287.

Papagno, C., Valentine, T., & Baddeley, A. D. (1991). Phonological short-term memory and foreign language learning. *Journal of Memory and Language*, *30*, 331-347.

Paradis, M. (2009). *Declarative and procedural determinants of second languages*. Amsterdam: John Benjamins Publishing Company.

Paribakht, T. S., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*(1), 9-29.

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

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 174-200). New York, NY: Cambridge University Press.

Paribakht, T. S., & Wesche, M. (1999). Reading and "incidental" vocabulary acquisition. An introspective study of lexical inferencing. *Studies in Second Language Acquisition*, *21*, 195-224.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*, 187-193.

Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*, 73-89.

Pigada, M, & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1-28.

Pilgreen, J. (2000). *The SSR handbook: How to organize and maintain a sustained silent reading program*. Portsmouth, NH: Heinemann.

Pimsleur, P. (1967). A memory schedule. *Modern Language Journal*, *51*, 73-75.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Polanyi, M. (1962). *Personal knowledge: Toward a post-critical philosophy*. New York, NY: Harper & Row.

Posner, M. I., & Rothbart, M. K. (1992). Attention and conscious experience. In A. D. Milner & M. D. Rugg (Eds.) *The neuropsychology of consciousness* (pp. 91-112). London: Academic.

Posner, M. I., Walker, J. A., Friedrich, F. J., & Rafal, R. D. (1987). How do the parietal lobes direct covert attention? *Neuropsychologia*, *25*(A), 135-146.

Posner, M. N. (2012). *Attention in a social world*. New York: Oxford University Press.

Pulido, D. (2004). The effect of cultural familiarity on incidental vocabulary acquisition through reading. *The Reading Matrix*, *4*(2), 20-53.

Qian, D. D. (1999). Assessing the roles of depth of breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*, 282-308.

Quine, W. V. (1960). *Word and object*. Cambridge, MA: MIT Press.

Rankin, E. F., & Overholser, B. M. (1969). Reaction of intermediate grade children to contextual clues. *Journal of Reading Behaviour, 1*, 50-73.

Rasinski, T. V. (2004). *Assessing reading fluency*. Honolulu, HI: Pacific Resources for Education and Learning.

Rea, C. P., & Modigliani, V. (1987). The spacing effect in 4- to 9-year-old children. *Memory and Cognition*, *15*, 436-443.

Readence, J. E., Bean, T. W., & Baldwin, R. S. (2004). *Content area literacy: An integrated approach*. Dubuque, IA: Kendall/Hunt.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355-371.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics*, *24*, 146-161.

Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology*, *55*, 297-308.

Rice, G. E. (1986). The everyday activities of adults: Implications for prose recall-part 1. *Educational Gerontology*, *12*, 173-186.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*(1), 77.

Rieder, A. (2003). Implicit and explicit learning in incidental vocabulary acquisition. *Views, 12*, 24-39.

Rinehart, S. D., Stahl, S. A., & Erickson, L. G. (*1986*). *Some effects of summarization* training on reading and studying. *Reading Research Quarterly*, *21*, 422-438.

Ringbom, H. (1978). The influence of the mother tongue on the translation of lexical items. *Interlanguage Studies Bulletin*, *3*(1), 88-100.

Ringbom, H. (1986). Crosslinguistic influence and the foreign language learning process. In M. Sharwood Smith & E. Kellerman (Eds.), *Crosslinguistic influence in second language acquisition* (pp. 150-162). Oxford, UK: Pergamon Press.

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Philadelphia, PA: Multilingual Matters Ltd.

Roberts, T. S. (2006). The use of multiple choice tests for formative and summative assessment. Proceedings of the 8th Australian Conference on Computing Education, Hobart, Australia. 175-180.

Rodrigo, V. (2006). The amount of input matters: Incidental acquisition of grammar through listening and reading. *The International Journal of Foreign Language Teaching*, *2*(1), 10-13.

Roediger, H. L., & Gallo, D. A. (2001). Levels of processing: Some unanswered questions. In M. Naveh-Benjamin, M. Moscovitch, & H. L. Roediger (Eds.), *Perspectives on human memory and cognitive aging: Essays in honour of Fergus Craik* (pp. 28-47). New York: Psychology Press.

Rohrer, D., & Taylor, K. (2006). The effects of over learning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*, 1209-1224.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition through reading. *Studies in Second Language Acquisition*, *21*, 589-619.

Rott, S., & Williams, J. (2003). Making form–meaning connections while reading: A qualitative analysis of word processing. *Reading in a Foreign Language*, *15*(1), 45-75.

Rydland, V. (2012). How word decoding, vocabulary and prior topic knowledge predict reading comprehension. A study of language-minority students in Norwegian fifth grade classrooms. *Reading and Writing*, *25*(2), 465-482.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*(3), 343-360.

Saragi, P., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*, 72-78.

Sarter, M., & Lustig, C. (2009). Attention and learning and memory. *New encyclopedia of neuroscience.* Oxford: Elsevier.

Schacter, D. L. (2001). The seven sins of memory: How the mind forgets and remembers. Boston, MA: Houghton, Mifflin.

Schmidt, R. W. (1995). Consciousness and foreign language: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning and teaching* (pp. 1-63). Honolulu, HA: University of Honolulu.

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences*, *28*, 763-774.

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and Second Language* (pp. 3-32). Cambridge, MA: Cambridge University Press.

Schmidt, R.W. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker (Eds.), *Proceedings of CLASIC 2010* (pp. 721-737). Singapore: National University of Singapore.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp.199-228). Cambridge, MA: Cambridge University Press.

Schmitt, N., & Schmitt, D. (1993). Identifying and assessing vocabulary learning strategies. *Thai TESOL Bulletin*, *5*(4), 27-33.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (l956). *A Study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223-250). Hillsdale, NJ: LEA.

Scott, J. A. (2005). Creating opportunities to acquire new word meanings from text. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary* (pp. 69-91). Mahwah, NJ: Lawrence Erlbaum.

Scrivener, J. (1994). *Learning teaching. A guidebook for English language teachers*. Oxford: The Bath Press.

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2004). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*, 107-122.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*,*10*, 209-31.

Sellers, V. D. (2000). Anxiety and reading comprehension in Spanish as a foreign language. *Foreign Language Annals*, *33*(5), 512-521.

Shettleworth, S. J. (1972). Constraints on learning. In D. S. Lehrman, R. A. Hinde, & E. Shaw (Eds.), *Advances in the study of behavior*. New York: Academic Press.

Shipley, E. F., & Shepperson, B. (1990). Countable entities: Developmental changes. *Cognition*, *34*(2), 109-136.

Shiwu, W. (2005). Use and helpfulness rankings of vocabulary learning strategies employed by EFL learners in Taiwan. *Journal of Humanities and Social Sciences*, *1*(2), 7-13.

Shokouhi, H., & Maniati, M. (2009). Learners' incidental vocabulary acquisition: A case on narrative and expository texts. *English Language Teaching*, *2*(1), 13-23.

Siegel, J. (2010). *Second dialect acquisition*. Cambridge: Cambridge University Press.

Singleton, D. (2001). Age and second-language acquisition. *Annual Review of Applied Linguistics*, *21*, 77-89.

Siriwan, M. (2007). *English vocabulary learning strategies employed by Rajabhat university students* (Doctoral Dissertation in English language studies, Graduate School, University of Technology).

Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology*, *111*(1), 65-86.

Smirnov, A. A. (1973). *Problems of the psychology of memory*. New York, NY: Plenum.

Smith F. (1971). Understanding reading: A psycholinguistic analysis of reading and learning to read. New York, NY: Holt, Rinehart & Winston.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568.

Smith, B. L., Smith, T. D., & Hobby, M. (2005). Relationship between intelligence and vocabulary. *Perceptual Motor Skills*, *100*(1), 101-108.

Smith, F. (1975). *Comprehension and learning.* New York, NY: Holt, Rinehart.

Smith, F. (1994). *Understanding reading* (5th ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Snowling, M., & Hulme, C. (2008). *The science of reading: A handbook.* Wiley.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*, 763-767.

Sprouse, R. A. (2006). Full transfer and relexification: Second language acquisition and creole genesis. In C. Jourdain, C. Lefebvre, & L. White (Eds.), *Montreal dialogues: Processes in L2 acquisition and Creole genesis* (pp.169-183). Amsterdam: John Benjamins.

Sternberg, R., & Powell, W. (1983). Comprehending verbal comprehension. *American Psychologist*, *38*, 878-893.

Stokes, J., Krashen, S., & Kartchner, J. (1998). Factors in the acquisition of the present subjunctive in Spanish: The role of reading and study. *Review of Applied Linguistics*, 121-122, 19-25.

Stringer, D. (2008). What else transfers? In R. Slabakova (Ed.), *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference* (pp. 233-241). Somerville, MA: Cascadilla Proceedings Project.

Swain, M., & Lapkin, S. (2005). The evolving socio-political context of immersion education in Canada: Some implications for program development. *International Journal of Applied Linguistics*, *15*, 169-186.

Swan, M. (2006). The influence of the mother tongue on second language vocabulary acquisition and use. In N. Schmitt & M. McCarthy (Eds), *Vocabulary: Description, acquisition and pedagogy* (pp. 156-180). Cambridge: Cambridge University Press.

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*, 261-285.

Tang, J. (2002). Using L1 in the English classroom. *English Teaching Forum*, *40*(1), 36-42.

Tavakoli, E., & Kheirzadeh, K. (2011). The effect of font size on reading comprehension skills: Scanning for key words and reading for general idea. *Theory and Practice in Language Studies*, *1*, 915-919.

Thalheimer, W. (2006). *Spacing learning events over time: What the research says.* Retrieved from http://www.work-learning.com/catalog/.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Reading Research Quarterly*, *21*, 422-438.

Trelease, J. (2006). *The read-aloud handbook.* New York, NY: Penguin Books.

Truscott, J. (1998). Noticing in second language acquisition: A critical review. *Second Language Research*, *14*(2), 103-135.

Ullman, M.T., Corkin, S., Coppola, M., Hickok, G., Growden, J.H., Koreshetz, W.J., and Pinker, S. A. (1997). Neural Dissociation within Language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, 9, 266-276.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition*, *4*,105-122.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*, 231-270.

Ullman, M.T. (2005). A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. In Sanz, C. (Ed.), *Mind and context in adult second language acquisition: Methods, theory, and practice* (pp. 141–178). Washington, DC: Georgetown University Press.

VanDeweghe, R. (2008). Research matters: Authentic literacy and student achievement. *The English Journal*, *97*(6), 105-108.

Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences*, *14*(4), 651-669.

Vlach, H. A., Sandhoffer, C., & Kornell, N. (2008). The spacing effect in children's memory and category induction. Cognition, 109, 163-167.

Vlach, H. A., & Sandhoffer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development*, *83*(4), 1137-44.

Waring, R. (1999). *Tasks for assessing receptive and productive second language vocabulary* (Unpublished doctoral dissertation, University of Wales, Swansea).

Waring, R., & Nation, P. (2004). Second language reading and incidental vocabulary learning. *Angles on the English Speaking World*, *4*, 97-110.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 130-163.

Waring, R. (2000). The State Rating Task—An alternative method of assessing receptive and productive vocabulary. *Studies in Foreign Languages and Literature*, *24*(1), 125-154.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46-65.

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232-245.

Werker, J.F., & Tees, R. C. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavioural Development*, *7*,49–63.

Wesche, M., & Paribakht, S. (1996). Assessing vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review, 53*(1), 13-40.

Williams, M. (1994). Motivation in foreign and second language learning: An interactive perspective. *Educational and Child Psychology, 11*(2), 77-84.

Williams, J.N. (2005). Learning without awareness. *Studies in Second Language Acquisition*,*11*,269-304

Willis, D. (1990). *The lexical syllabus.* Winston: Harper Collins.

Wilson, K. V. (2011). *From associations to structure.* New York, NY: North-Holland Publishing.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and the containing function of wrong beliefs in young children's understanding of deception.*Cognition*,*13*, 103-128

Yaqubi, B., Rayati, R. A., & Allemzade, G. (2010). The involvement load hypothesis and vocabulary learning: The effect of task types and involvement index on L2 vocabulary acquisition. *Journal of Teaching Language Skills*, *1*, 145-163.

Yoshida, M. (2007). *Think-aloud protocols and type of reading task: The issue of reactivity in L2 reading research*. Kyoto University of Foreign Studies, Japan: Kyoto University Press.

Zimmerman, C. B. (1997). Do reading and interactive vocabulary instruction make a difference? An empirical study. *TESOL Quarterly, 31*(1), 121-140.