BRITISH COUNCIL

UNIVERSITY OF BIRMINGHAM

Measuring Vocabulary Knowledge Growth with a State Rating Task Self-Reporting Instrument

by Alex Wright

British Council's Master's Dissertation Awards 2022

Special Commendation

Measuring Vocabulary Knowledge Growth with a State Rating Task Self-Reporting Instrument

by

Alex Wright

A dissertation submitted to the College of Arts and Law of
the University of Birmingham in part fulfilment of the
requirements for the degree of

Master of Arts

in

Teaching English to Speakers of Other Languages (TESOL)

This dissertation consists of approximately 15,000 words.

Supervisor: Theron Muller

English Language and Linguistics
College of Arts and Law
University of Birmingham
Edgbaston, Birmingham B15 2TT
United Kingdom

September 2021

Abstract

This dissertation describes an investigation into the differences in receptive and productive vocabulary knowledge growth of Japanese university English department students through a self-reporting instrument. One group was made with students of a higher level extensive reading class and a second group was made with students from a lower level extensive reading class. Both groups undertook the same standard extensive reading curriculum for 1 semester. The higher level class also received explicit weekly vocabulary lesson videos. A State Rating Task questionnaire containing target words that appeared in the graded reader library and explicit lessons was constructed, and 4-states of knowledge were provided on a scale for students to self-report. Both groups took the same questionnaire at the beginning and end of the semester. The higher level group reported significantly higher levels of productive vocabulary knowledge than the lower level group at the end of the semester. There did not appear to be any significant difference in reported receptive vocabulary knowledge. The findings of this investigation were consistent with previous findings that a mixture of incidental and intentional input is more effective for vocabulary knowledge growth than strictly incidental input.

Acknowledgements

I would like to express my sincerest gratitude to Theron Muller for his guidance throughout the composition of this dissertation, and to the student volunteers who took part in the study. I would also like to thank Patrick Conaway for offering helpful advice and Tomomi Yoshida for assisting with construction of the Japanese questionnaires. Finally, I owe thanks to John Wiltshier for helping with the design of the investigation.

Table of Contents

List of Tables and Figures

CHAPTER 1
INTRODUCTION

The idea for this dissertation first materialized when a co-worker asked me if it is truly necessary for us to spend time teaching English words that frequently appear in English books that students read as part of their extensive reading program. This became a question that I wanted to explore and test for my own classes, and I knew that a method for answering this question most likely resided in the extensive amount of research on the effectiveness of reading as educational input for learning. I came across several popular publications making approximations such as those estimating that the chance of learning a new word from context is around 5 percent (Nagy et al., 1987) or the optimal inclusion of unknown words in a text for comprehension and learning is around 2 percent (Coady & Nation, 1988). Numbers like these do not seem terribly efficient, but I noticed that the definition of "learning" in these publications seems to refer to going from a near zero state of knowledge to a state of usability or mastery, and likely excludes most of the journey between these states. Fairly confident that knowledge in the human mind is not a binary mechanism in this way, I began to investigate to what extent researchers had taken a deeper look at the levels of knowledge that language learners might be progressing through when learning.

This investigation revealed that there are in fact theories about the types of scales that knowledge progresses along. Henriksen (1999) describes how we can know vocabulary to different levels of precision, in different contexts, or for different uses in comprehension or production of language. The idea of the

existence of these different scales is intriguing, but the tools to measure them do not seem to be as straightforward as normal tests. For example, we can ask how a subject may provide the answer to a fill in the blank question when they have some idea of the answer but cannot recall the word exactly. Further, how would we rate the score of such answers if they were possible? One attempt to do this comes in the form of more qualitative instruments called vocabulary knowledge scales. These are often formed as Likert scale style self-reporting questions which may also provide opportunities to guess and provide synonyms or descriptions when subjects do not know an answer perfectly. An early and clear example is the 5-level Vocabulary Knowledge Scale (VKS) created by Paribakht and Wesche (1993).

One of the most comprehensive collections, comparisons, and criticisms of these scales came from Waring (1999). His doctoral dissertation explored how to clearly define and measure receptive vocabulary knowledge, the degree to which a subject can understand certain vocabulary, and productive vocabulary knowledge, the degree to which a subject can produce language with certain vocabulary. He acknowledged the problem that standard testing cannot easily measure levels of these types of knowledge and that many typical questions test not only for receptive-productive vocabulary knowledge, but for a mix of skills and knowledge that is difficult to separate. In this way, standard testing is usually not exclusively testing for specific knowledge types. This led him to self-reporting instruments, and he made convincing arguments that most of the recently created vocabulary knowledge scales are flawed in several ways. They may not clearly define language such as “use” or “meaning” that can be

understood as different things to different subjects, or they may skew the number of response options towards different types of knowledge. It is also the case that the people piloting the instruments often assign a linear interval scale to the scoring method, and perform descriptive or inferential statistics on the numbers of this scale. These scales are often clearly non-linear, which necessitates that we be skeptical about the conclusions drawn from the results of these analyses.

Waring (1999) attempted to rectify many of these flaws by constructing a self-reporting instrument and clearly defining "understand" and "use" on 2 different scales for his subjects. He created a scale from 0 to 3 that appeared linear to the subjects, but he did not assume the data was on a linear scale. Instead of actions like averaging the scores of a group, he looked at how many subjects were in certain states and how they changed over time for data analysis. The name that Waring (1999) gave to this type of instrument was the State Rating Task (SRT). He showed the applications of the idea with several experiments, and although the best method of data analysis appeared to be unclear, the data seemed to produce information that was clearly valuable to answering his research questions while remaining conservative with regards to unproven assumptions.

The idea of the SRT does not appear to have taken off, as there are only a few published studies which have used the instrument since 1999, but after considering the valid criticisms of Waring (1999) which are described in detail in this dissertation's review of the literature, I decided that a similar SRT instrument would be the best way to perform an more sensitive investigation on the

vocabulary knowledge of my own students and examine whether intentional teaching provided a significant advantage in learning over simply reading books.

The following chapters provide an overview of the relevant literature behind the topics above, and details about an experiment performed over a single 4-month university semester to answer the 3 questions:

1. What differences in student vocabulary knowledge change are there between an extensive reading course with and another similar course without supplementary vocabulary lessons?

2. How can extensive reading instructors set up their courses and materials to optimize vocabulary knowledge growth?

3. What are the advantages and disadvantages of the SRT as a research instrument to measure vocabulary knowledge?

After presenting the results, the degree to which these questions were answered and the the limitations that had the greatest impact on the investigation are discussed.

CHAPTER 2

LITERATURE REVIEW OF VOCABULARY KNOWLEDGE AND EXTENSIVE READING

The main goal of this dissertation is to determine the optimal conditions for improving vocabulary knowledge growth in subjects who are participating in extensive reading (ER). The instrument used to measure this improvement was a self-reporting instrument called the State Rating Task (SRT). In this literature review, views and research results on the nature of vocabulary, ER as a language learning tool, and several studies that bring these topics together in attempts to measure vocabulary knowledge growth through ER are discussed. The concept of the SRT and the characteristics that make it unique among self-reporting vocabulary knowledge scales are also presented. These topics are central to understanding the topics and results of this dissertation.

2.1 Vocabulary in Language Learning

As this dissertation is primarily concerned with vocabulary knowledge, it is important to understand where we stand when it comes to our understanding of vocabulary. There have been a wide variety of evolving views on the place of vocabulary in language education, as well as how and to what extent learners' knowledge may grow with certain treatments. Vocabulary, or more formally the lexis of a given language, refers to the set of all words that are used. The word is arguably the most fundamental unit of human languages, and is thus of utmost importance for language learners to learn. Smaller units exist, such as morphemes, and larger units convey even more complex relationships, such as clauses, but the

unit of a single word on a page often permits a complete image of an object or concept. Wilkins (1974: 111-112) wrote, "…without grammar very little can be conveyed, without vocabulary nothing can be conveyed." Further, the definition of a word is not a consensus among scholars (Haspelmath, 2011), but in this dissertation a word is defined as a single written English unit with no separating spaces.

The following subsections discuss popular modern views within the language teaching community on vocabulary in education, intentional and incidental learning, knowledge continua, self-reporting instruments, and the SRT, as they relate to vocabulary knowledge.

2.1.1 Views on the Place of Vocabulary in Language Learning

A symbiotic relationship between vocabulary and language such that having vocabulary enables language use, and using language will lead to improvements in vocabulary knowledge was described by Nation (2001). Nation (2001) also deepened our considerations of vocabulary as having the elements of *form, meaning,* and *use*. *Form* is described as the physical manifestation of the word, whether that be written or spoken. *Meaning* is described as the ideas that the words represent, and how they can be explained using other words. Finally, *use* is described as the patterns that emerge when the word is actually used, and the contexts in which it appears. While these classifications were helpful to consider words from different perspectives, the question of which words are the most essential for learners to master was still unclear. One body of work that helped to concretely answer this question was that of determining word frequency. The

Collins COBUILD corpus was a beginning of such work (Sinclair, 1987), but it became even more refined with collections such as the New General Service List (NGSL) and the combination of the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) (Nation, 2016), which was used in this dissertation. These lists organize vocabulary into groups by the frequency at which they occur across the English language and have helped create tests and study plans which allow learners to gauge where they should focus their efforts (Nation, 2016). Considering frequency has also proven useful when designing the instruments used in this dissertation.

2.1.2 Intentional and Incidental Vocabulary Acquisition

One key debate that this dissertation engages in is whether vocabulary is acquired more intentionally from deliberate study, or incidentally through the context of natural input. This dichotomy in models of language learning is also described as explicit and implicit (Ellis, 1994), active and passive (Laufer, 1998), and decontextualized and contextualized (Oxford & Crookall, 1990). Graves (2000) describes intentional vocabulary learning as involving explicit teaching of words, vocabulary learning strategies, and activities in which subjects engage in consciously learning the meanings and usage of words. On the contrary, incidental learning occurs when subjects use a language and begin to infer the meaning of new vocabulary from context. Incidental learning of vocabulary can occur during ER, conversations, or listening to the language being spoken (Nation, 2001). Other factors are known to affect the rates of these types of learning, such as the frequency of a given target word across all human discourse and the comprehensibility of its surrounding language context.

Several prominent scholars have expressed their belief that incidental vocabulary learning has a greater role (Krashen, 1989; Nagy, 1997); however, they acknowledge that for true gains to be made, massive amounts of input (i.e., ER) are required to introduce and reinforce vocabulary in the minds of learners. Nagy et al. (1987) found that the chance of learning a new word from context through reading was only around 5%. Most findings seem to agree that mass amounts of reading will lead to slow and steady gains for more advanced learners, yet in order to begin free-form input activities such as ER, a certain vocabulary minimum must be obtained. Thus, especially when beginning to learn a language, intentional learning is far more efficient (Huckin & Coady; 1999, James, 1996) and is a prerequisite to entering the field of incidental learning (Schmitt, 2000).

As with many competing pedagogic theories in the educational world, we may find that a balanced approach rather than an extreme adherence to one type of learning is the most effective for our desired endpoint. Waring & Nation (2004) explain this synergy with regards to intentional and incidental learning:

> "All studies comparing incidental with intentional learning show that intentional learning is more efficient and effective. This should not be seen as a competition between incidental and intentional learning. Rather, a well balanced language programme should make good use of both types of learning. One without the other is inadequate." (Waring & Nation, 2004: 20)

There are various studies that support this idea of combining the two types of learning, and in particular see Knight (1994), Meganathan et al. (2019), and Zahar et al. (2001) for examples. This dissertation also tests the benefits of a combination of intentional and incidental learning, versus more strictly incidental

learning. Consequently, it will prove useful to compare the conclusions of this dissertation with those above.

The species and context of words affect the discussion of which type of learning may be more effective. Nation (2005) argues that low-frequency words are better suited for incidental learning, once the necessary foundation of frequently occuring words has been acquired. However, this can lead to an unfortunate cycle because low-frequency words do just that—appear infrequently (Nation, 1990). This means less interactions and less chances for learning, unless the volume of words read is further increased. It is also estimated that for subjects to comprehend a written text, the percentage of unknown words must be downwards of 5% (Hirsh & Nation, 1992; Laufer, 1989; Nation, 2009). All of the information to this point suggests that the intermediate level university student subjects of this dissertation acquire a significant portion of new and rare vocabulary through incidental learning, and frequent input through sources such as ER would be one of the best methods to facilitate this growth.

2.1.3 Vocabulary Knowledge Continua

Continuous scales upon which vocabulary knowledge levels reside may exist, but they do not appear to be well supported by evidence such as testing. They may be exclusive or overlap with each other, but it remains very difficult to isolate and test such propositions. This dissertation attempts to use some of these continua while at the same time investigating the claims about them and assessing whether the theories are consistent with the evidence that is collected.

3 dimensions in which we may think about vocabulary knowledge have been described (Henriksen, 1999). The first is a partial-precise dimension where precise knowledge allows the completion of difficult tasks such as translation or explaining the meaning of a word using other words. In contrast, partial knowledge refers to a vaguer awareness of a word and its meaning. The second dimension is the depth of knowledge scale which refers to knowledge about a word's multiple meanings in different contexts, such as what kinds of collocations are common. The third dimension is the receptive-productive dimension which relates more to the ability to actually comprehend a word during reading and listening versus the ability to produce something through speaking and writing.

Henriksen suggested that we may be able to measure knowledge along the partial-precise continuum using a series of tasks (Haastrup & Henriksen, 1998), and because low knowledge would only allow completion of easier tasks it was thought that a difficulty scale would directly relate to a knowledge scale. However, the results did not appear to show clear progression in score from easy to difficult tasks. Waring (1999) proposed that this is because there were simply too many different aspects to the tasks in Haastrup & Henriksen's (1998) study when it came to the knowledge and skills required to complete them (i.e., reading, speaking, sorting). This demonstrates the difficulty in designing instruments to measure such theories of vocabulary knowledge.

Receptive and productive vocabulary (RPV) knowledge have also been investigated in similar ways. Henriksen (1999) argued that the RPV knowledge

dimensions lie on the same continuous scale and need not be separated, yet they often are. One of the earliest separations and descriptions of RPV knowledge was Gaultier (1839) in Kelly (1969), who described a passive type of vocabulary knowledge as facilitating understanding and an active type which allows subjects to compose. Assuming this distinction exists, this dimension is likely related to basic retrieval cognitive processes such as recognition (recognizing an item when seen) and recall (recalling an item from memory with minimal prompting, if any) (Wolfe, 1886). Later on, these types of ideas in relation to vocabulary were more concretely measured by Myers (1914) who found that subjects' abilities to recognize were over twice as good as their abilities to recall in simple vocabulary tests and follow-up assessments. These tests involved recognition, through identifying a word among various incorrect answers, and recall, through writing a word that had been previously seen from memory.

Over the years, the definitions of these classifications of vocabulary knowledge and psychological processes seemed to grow and encompass a large area of topics. One possible source of ambiguity in how studies have approached and reported investigations of RPV knowledge is that the definitions of these phenomena became fluid and interchangeable. Waring draws attention to this in his doctoral dissertation:

> "The problem we are now faced with is that the simple psychological notions of recall and recognition became common parlance within educational psychology without so much as a recognition that the terms used in psychology were not necessarily appropriate. A recognition test in psychology is not necessarily a test of the receptive skill of reading, nor is a recall test a test of the productive skill of writing. This confusion still besets us." (Waring, 1999: 4)

Waring (1999) argues that because of these fluid definitions there are discrepancies between tests that are labelled with one of these cognitive process keywords or vocabulary knowledge keywords and the actual types of knowledge that are required to perform them. For example, a test utilizing recognition-type multiple choice questions that is labelled "receptive" is not exclusively measuring receptive vocabulary knowledge, as it may include other mental processes such as comparing or skills that might be used in common strategies like process of elimination. For the purposes of this dissertation, the definitions of RPV knowledge are considered to be the ability to make sense of vocabulary for reading and listening comprehension (receptive), and the ability to use vocabulary for language production in writing and speaking (productive). The data collection instrument used in this dissertation to acquire data also attempts to step outside the complex requirements of other common vocabulary test instruments and measure this RPV knowledge more directly.

2.1.4 Self-reporting Vocabulary Knowledge Instruments

New instruments for testing vocabulary knowledge have arisen due to the questionable validity in testing described in Section 2.1.3. Arguing that the layer of tasks involved in the process of testing vocabulary knowledge may be an unnecessary layer of assumptions, Waring (1999: 58) states, "The simplest way for a researcher who is interested in finding out if a learner knows a word, is to ask her." This brings the instrument into the field of self-reporting, which may be more direct but comes with its own limitations. There is an assumption that subjects are aware of their own knowledge and able to report it. Questioning the

validity of this assumption begins to enter the realm of psychology more than language teaching, but there is some evidence that learners can reliably do so.

Schouten-Van Parreren (1996) set up a study where learners were asked to self-report their knowledge, and then asked to perform a comprehension test. She found that students tended to overrate their knowledge of certain words, such as those that were similar to words in their first language but actually had a different meaning in the target language. Even including such discrepancies, she found a correlation between the self-reports and the actual test scores of 0.59, which is a somewhat strong correlation. The reliability of such self-reporting instruments can be further improved by adding controls, such as pseudo-words that do not exist or words that are expected to be in a state of near complete knowledge. Subjects who do not report these words as they are expected to can be considered unreliable and removed from the data pool. In testing that is aiming to observe relative changes rather than absolute changes (e.g., for research as opposed to standardized testing), some self-reporting inaccuracy may not pose an issue, as long as it is consistent within the individual (Waring, 1999).

If we assume that an acceptable number of reliable answers can be obtained, we must then ask what the questions and answers on a self-reporting instrument would look like. Many instruments have been created and their wordings vary considerably. The simplest example may be a YES or NO checklist where the word is reported as known or unknown, such as that used by Law (1991). While this type of checklist may be useful for surface level surveys, greater depth will often be required if we desire more sensitivity in observing progression. A type of

scale would then seem appropriate to see finer details in knowledge levels. Table 2-1 shows 3 such scales through which subjects could report their knowledge through multiple choice questionnaires.

Zimmerman (1997)
(How well do you know this word?) 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 when I see it or hear it in a sentence, but I do not use it in my own speaking or writing. D) I can use the word in a sentence.
Wesche and Paribakht (1996)
(How well do you know this word?) I) I don't remember having seen this word before. II) I have seen this word before, but I don't know what it means. III) I have seen this word before, and I think it means_____ (synonym or translation). IV) I know this word. It means_____ (synonym or translation). V) I can use this word in a sentence. E.g.:_____ (If you do this section please also do section IV)
Tan et al. (2016)
(How well do you know this word?) 1) I do not think I have ever seen this word.__ (Tick if true, do not proceed) 2) I have seen this word, but I do not know what it means.__ (Tick if true, do not proceed) 3) I have seen this word before and I think it is related to the following word/idea_____(answer may be given in English/Baha Malaysia). 4) I have seen this word before and I think it means:_____(give a synonym in English/Baha Malaysia) 5) I cannot use this word in a sentence. __ (Tick if true, do not proceed) 6) I can use this word in a sentence:_____(write your sentence in English) Translate your sentence in Baha Malaysia:_____.

Table 2-1. 3 different vocabulary knowledge scales.

In Table 2-1 we can see various levels of depth and verification. Tan et al.'s (2016) questionnaire shows 7 levels of potential progression and demonstration, while Zimmerman's (1997) questionnaire simply has 4 multiple choice options. Wesche and Paribakht's (1996) questionnaire is somewhere in the middle.

Interestingly, all of these scales attempt to give the subject a score by assigning a value to their answers, and also attempt to compare them on a linear scale. For example, Wesche and Paribakht (1996) assigned a numerical score for each word equal to each level of the scale from 1-5. Further, a failure to demonstrate correctly in levels III, IV, and V by the subject resulted in a regression to a lower score. The original authors may not have intended for this scoring system to be used as an interval scale in descriptive and inferential statistics involving averages, but it has been in some instances (Pulido, 2004), leading some to question if this numerical scoring scheme actually has any value in data analysis (Bruton, 2009). The numerical scoring systems that Zimmerman (1997) and Tan et al. (2016) assigned to the levels of their scales were also used in statistical comparisons of the mean score.

There are several problems in the design and scoring of these scales. First, verification may seem desirable, but the verifications that are included transform these questionnaires into a form of test with similar problems to those discussed in Section 2.1.3. Specifically, answering these questions requires skills such as writing or recalling synonyms, and if we are trying to measure something specific, such as receptive or productive vocabulary knowledge, there are now extra layers of assumptions. Wording is also unclear. For example, in Wesche and Paribakht's (1993) questionnaire, does "use" mean in speaking or in writing, or both? Furthermore, it hardly seems correct for us to be able to assign each of these levels a numerical score before collecting the data, or place them on a linear scale. Using Tan et al.'s (2016) questionnaire as an example, who are we to say that knowing a vague meaning of a word (Level 3, worth 2 points) is worth twice as

much as simply having seen it before (Level 2, worth 1 point)? What does an average score of 1.5 actually mean? If a linear scale was the correct assumption, a 1.5 would mean something along the lines of "The subject has seen the word, but is halfway between the situations described in Levels 2 and 3", which becomes very difficult to describe concretely. A potential solution to these issues is discussed in Section 2.1.5.

2.1.5 The State Rating Task

Waring (1999) also realized several of these issues, and came up with a potential solution. He describes a type of multi-state model of vocabulary knowledge development called the State Rating Task (SRT). The multi-state model views vocabulary knowledge as existing with certain attributes at a given point in time. The change between states is what can then be analyzed. For the basis of this idea, Waring (1999) cites the matrix analysis and probability state models of Meara & Rodriguez Sanchez (1993), which is not currently available, but see Meara & Rodriguez Sanchez (2001) or Meara (1996) for similar studies. A visualization of vocabulary shifting in SRT states can be seen in Figure 2-1 below.

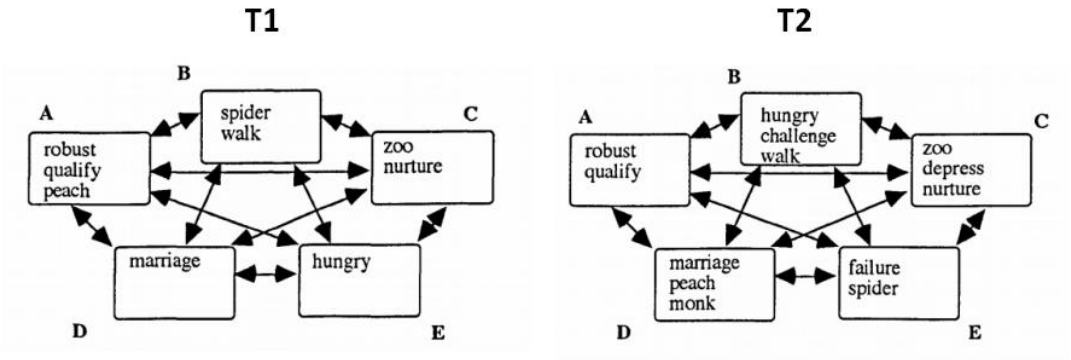

Figure 2-1. A multi-state model for vocabulary knowledge, changing from Time 1 (T1) to Time 2 (T2) (Waring, 1999).

In Figure 2-1, we can imagine that any of the letters A to E reflect states with descriptions similar to Zimmerman's (1997) vocabulary knowledge scale shown in Table 2-1. In contrast to a vocabulary knowledge scale, the multi-state model implies that words in any state can move to words in any other state, and there is no immediate mathematical linear progression. These states should initially be treated as nominal, meaning they have no numerical value, but rather we are observing how changes proceed over time. In order for the states to be treated as ordinal, defined as each state representing a clearly higher or lower level of knowledge than each other state, the descriptors need to be simpler and clearer such that there is no mixture of skills or scales, or ambiguous definitions.

Visualizing vocabulary knowledge states in diagrams like Figure 2-1 could be useful on its own, but to compare large groups of subjects, another form of data analysis is desirable. The immediate data does not provide the usual mean or

median that one might find statistically useful, but the changes can be portrayed in a matrix.

T2

T1	(→)State 0	(→)State I	(→)State II	(→)State III
State 0	15	3	0	1
State I	4	20	6	9
State II	1	3	14	5
State III	0	2	4	12

Table 2-2. Hypothetical T1 to T2 state changes of 99 words from an individual or group (Waring, 1999).

In Table 2-2 we can see a hypothetical example of a number of words that have shifted into new states given an initial starting state, or stayed static. Values on the diagonal represent the numbers of words that have not shifted, but rather remained static in their original state (shown in gray). We might expect most words not to shift over short time periods, but we might see State I(→)State III, and consider it an abnormally high shift of 9 words. If we decide that State III is highly desirable knowledge state then we can begin to make other connections. For example, looking at the real life situation may reveal a reasonable explanation for the high shift, such as more exposure to such words.

Such data analysis methods have not been highly refined to accommodate for various research goals as there appear to be only 2 studies using the SRT, but those that were used are described here. In his Experiment 2, Waring (1999: 84) used several similar matrices to show how 2 different level groups differed in

vocabulary knowledge. They first performed a spew task, in which subjects had to list as many words as they could with certain parameters. One matrix showed states of time versus the number of words produced, and a second matrix was also produced to show the frequency band of words versus the number produced. Yet a third matrix involved an SRT in which students rated how sure they were that they would recognize a given produced word the next time they saw it. The results of this SRT were then again contrasted against the word frequency bands in a matrix. The scale of the SRT was from 0 to 3, providing 4 different states, and this data was treated as nominal according to Waring (1999: 73). The states were labelled *0: I am sure I will not know*, *1: I am not sure I will know*, *2: I am quite sure I will know*, and *3: I will definitely know*. The results of this study allowed Waring (1999) to perform some data analysis by treating states 1, 2, and 3 together as having some knowledge of a word, and treating state 0 as having no knowledge. He also corrected for students who rated pseudo-words highly. Waring then commented on the high Guttman scalability (Guttman, 1944) of the test procedure, which is a measure of how well a test can distinguish between levels of what it is attempting to measure. High scalability may allow us to infer a more accurate numerical scaling for the states, and to make statements such as: students who know a certain number of words in frequency band 5 will know all words in bands 1 to 4. Waring (1999) noticed the phenomenon that most words produced in the spew task were rated as being in a high state of knowledge, as well as noting a decrease of highly rated words in the SRT as one increases the frequency band.

Dabaghi & Rafiee (2012) also used an SRT instrument to attempt to measure vocabulary knowledge growth during reading with the assistance of first or second language glosses, which are helpful definitions along the margins of a text. The students self-reported their knowledge before and after the readings using the states described in Table 2-3 below.

E	D	C	B	A
I do not know the word.	I think I know the word, but do not know how to use it.	I think I know the word and how to use it.	I know the word, but do not know how to use it.	I know the word and how to use it.

Table 2-3. The SRT states used by Dabaghi & Rafiee (2012).

The states shown in Table 2-3 appear to start showing ordinal progression from undesirable to desirable, but it becomes difficult for us to judge whether state C or B is a more desirable state to be in, which shows why this rubric is more suitable for a nominal SRT analysis than traditional numerical scoring. Dabaghi & Rafiee (2012) also describe how students who answer D and B are assumed to have receptive vocabulary knowledge, but not productive. Those who answer C or A have have both. Seemingly just by observing patterns in the movement of words through states, such as considering the difference between same state, near state, and dramatic state changes, the SRT results found in this study led researchers to conclude that the subjects showed considerable growth in overall vocabulary knowledge through the activity, and that first language glosses were more effective at improving productive vocabulary while second language glosses were more effective at improving receptive vocabulary.

The conclusions described in this section appear to be valuable, but it may be more beneficial to make more use of all the different SRT states if they can be assumed to be on ordinal or linear scales. The literature to this point defines SRT states as nominal, but appears to use ordinal data properties to some degree by describing some states as more favorable than others. An argument can be made that Waring's (1999) states in the tasks above were more ordinal than nominal, as any given state is clearly described as representing higher or lower knowledge than the others. This dissertation has attempted to recreate a similar task and instrument, and justify the data being on a linear or ordinal scale. It then used traditional statistical methods to analyze the data as linear, as well as more novel methods to analyze the data as ordinal.

2.2 The Rise of Extensive Reading

Extensive reading (ER) is a central component to this dissertation`s investigation, and so it is necessary to clearly define ER and explain how it has become a valued part of many language programs around the world. The term extensive reading began seeing use in the 20th century as a juxtaposition to intensive reading, and is often credited to Palmer (1917). Bamford & Day (1997) describe ER as "generally associated with reading large amounts with the aim of getting an overall understanding of the material." Essentially, ER is associated with reading for pleasure, self-selecting books, reading freely over large periods of time, and reading for the greater meaning rather than that of individual words, or even sentences. Bamford & Day (1998) also clearly outlined 10 qualities of an effective ER program:

(1) *Students read as much as possible*, perhaps in and definitely out of the classroom.

(2) *A variety of materials on a wide range of topics is available* so as to encourage reading for different reasons and in different ways.

(3) *Students select what they want to read* and have the freedom to stop reading material that fails to interest them.

(4) *The purposes of reading are usually related to pleasure, information and general understanding*. These purposes are determined by the nature of the material and the interests of the student.

(5) *Reading is its own reward*. There are few or no follow-up exercises to be completed after reading.

(6) *Reading materials are well within the linguistic competence of the students* in terms of vocabulary and grammar. Dictionaries are rarely used while reading because the constant stopping to look up words makes fluent reading difficult.

(7) *Reading is individual and silent*, at the student's own pace, and, outside class, done when and where the student chooses.

(8) *Reading speed is usually faster rather than slower* as students read books and other material that they find easily understandable.

(9) *Teachers orient students to the goals of the program*, *explain* the methodology, *keep track* of what each student reads, and *guide* students in getting the most out of the program.

(10) *The teacher is a role model of a reader for students* -- an active member of the classroom reading community, demonstrating what it means to be a reader and the rewards of being a reader.

(Bamford & Day, 1998: 7-8)

There exist other descriptions of different types of effective ER programs (Waring & McLean, 2015), but the ER class subjects that participated in this dissertation attempted to follow the 10 guidelines above, with the exception that the students received grades for reaching department set goals as a type of reward.

Bamford & Day (1998) also listed a wide variety of over 10 studies that show gains achieved by students in ER programs in almost all areas. Some of the more

convincing of these are Elley & Mangubhai (1981), which showed gains in reading, listening, and writing proficiency, and Cho & Krashen (1994), which showed gains in reading, vocabulary, and oral skill proficiency. Many of these findings suggest that ER is not only beneficial to reading skills, but also benefits other areas of language learning. More recent studies have also found similar results. He (2014) found that groups that incorporated ER into their routine saw improvements in reading, listening, and overall language proficiency, although she found that the greatest gains were seen when ER was a supplementary element of the syllabus rather than the main method of learning, which may support the benefits of combination approaches discussed at the end of Section 2.1.2. Nakanishi (2015) performed a meta-analysis of 34 results from previous studies, and found medium positive effect sizes for both group contrasts (d=0.46), and pre-post contrasts (d=0.71). The results of the past 50 years seem to have convinced many instructors that ER is worth implementing into organizations which have the resources to set up such a program. Some evidence for this is the appearance of ER focused organizations across the world such as the Middle East and North Africa Extensive Reading Foundation, the Japan Extensive Reading Association, and the Indonesian Extensive Reading Foundation.

2.3 Studies on Extensive Reading for Vocabulary Building

It is now necessary to clarify exactly what has been found experimentally when it comes to vocabulary knowledge growth and ER. Several studies that have already been mentioned contain links between ER and vocabulary, such as Nagy's (1987) 5% incidental acquisition rate through reading. Furthermore, considering Bamford & Day's (1998) collection of research, it is clear that

several studies from this time period showed vocabulary knowledge growth due to reading. The first was Pitts et al. (1989) who showed that adult learners who read a portion of a single book could learn the meaning of slang words significantly better than those who had not read the text, through a vocabulary recognition multiple choice test. This is a good example of incidental vocabulary learning through reading, but it can hardly be considered extensive. Next, Hafiz & Tudor (1990) showed that a 6-week ER program provided significant gains to the base vocabulary used by learners in a productive writing task when compared to a group who had not done ER. Lai (1993) then showed that students who had read more during a 4-week ER program showed significant gains in reading comprehension tests that contained vocabulary recognition tasks compared to those who had read less. Finally, as was mentioned in Section 2.2, Cho & Krashen (1994) found a variety of benefits in a small-scale study with 4 learners, and specifically they found that in a productive oral definition task that the subjects were able to define between 56% and 80% of words that they had previously underlined as unknown upon the first encounter. Qualitative interviews also revealed that the subjects believed their vocabulary knowledge in relation to speaking and listening also improved. Through extrapolation of their data, Choe & Krashen (1994) also estimated vocabulary acquisition rates of 5000, 2,500, 1200, and 1000 words learned per 1 million words read for their 4 subjects respectively.

These studies show the general trend that different aspects of vocabulary are improved by ER, but also that a wide variety of instruments have been used to measure such gains. It is difficult to compare the results of the above studies with

the results of this dissertation because these studies mostly operate only at 2 levels of knowledge: “known” and “unknown”; however, this dissertation aims to observe more than 2 levels of knowledge, which may show growth for a higher proportion of words, albeit not necessarily to the level of complete mastery.

There have been several propositions that ER does promote new vocabulary acquisition at low rates, but that much of its desired effect on vocabulary knowledge is in strengthening the knowledge of words that are already known (Nation, 2001; Waring & Takaki, 2003). Despite this, such a value remains much more difficult to measure. More sensitive test battery, or perhaps self-reporting instruments could provide the next steps in this field. As an example, Waring & Takaki (2003) used 3 different forms of testing for measuring vocabulary knowledge gain through the reading of a single graded reader, which were (1) a word form recognition test, (2) a multiple choice test, and (3) a translation test. They found that for each of these respective tests, on average the subjects were able to correctly give the meaning for (1) 15.3, (2) 10.6, and (3) 4.6 out of 25 previously unknown pseudo-word substitutes. This again suggests that the type of test is heavily influential in determining the amount of vocabulary knowledge that can be gained from ER. More sensitive investigation could also help to identify words that subjects have some knowledge of, but cannot correctly answer on a test, or help measure subjects’ knowledge of words for which they do not feel comfortable enough to write YES on a vocabulary knowledge checklist. Horst (2005) explored this to some degree by measuring vocabulary knowledge growth attributed to ER through a YES/NO checklist style of self-reporting, followed by a vocabulary knowledge scale inquiry for words that were rated NO at the

beginning of the program. She discovered that for 18 out of 35 instances of inquiry subjects did in fact display partial or full knowledge of the words that they had previously labelled unknown. Some of these findings are in stark contrast to other findings previously discussed in this dissertation which suggest that only several thousand words can be learned per million words read, but this is likely due to the increased sensitivity of the instruments.

In more recent studies, further depth into the reading process has also been investigated. By splitting ER class learners into 3 different levels through standard vocabulary levels test (VLT) scores, Webb & Chang (2015a) showed that their highest level students had close to 60% vocabulary gains in simple definition matching pre-post-delayed-tests while their lower level group could barely exceed 35% gains. Suk (2017) performed a study on the growth of reading comprehension, reading rate, and vocabulary, in several reading class contexts and found that vocabulary gains were the highest of the attributes studied. This study used a target language to first language single word translation test format, and the authors stated that their subjects' vocabulary knowledge gains were higher than other similar studies, likely because Suk's (2017) subjects consistently read similar level graded readers and the tests were directly based on words that frequently occurred rather than a standardized VLT.

2.4 Summary

Throughout this literature review, the relevant history of vocabulary use, knowledge, and growth has been discussed and a wide variety of theories and testing instruments that have been used were considered. The review also touched

on promising newer research instruments that have neither been well explored nor well established, and several studies that indicate ER class implementation policies that may optimize vocabulary knowledge gain. With these things considered, this dissertation has attempted to construct an instrument based on the promising qualities of the SRT, and measure RPV growth differences in 2 ER program classes with differing degrees of intentional and incidental instruction. Using this SRT, subjects self-reported their ability to understand and use several words that they were likely to encounter in their graded readers or lessons, which formed the basis of the data analysis.

CHAPTER 3

RESEARCH METHODS: MEASURING VOCABULARY KNOWLEDGE THROUGH AN SRT

The objective of this chapter is to clearly present the subjects of the experiment, the experimental set up, and the methods of data analysis in such a way that the experiment could be repeated by another party, or compared with research done in similar contexts.

3.1 Participants

The participants form an integral part of any research involving language learning. The backgrounds and process of recruitment of the participants involved in this dissertation are described in this section. Ethical permission was obtained from the university to conduct the experiments described in this dissertation with student volunteers.

3.1.1 Participants' Backgrounds

First, I will describe the study participants; 25 second-year Japanese university student volunteers. These students belonged to the English department of a medium-sized private women's university and were studying English as a foreign language. As a result, many had some degree of interest in learning English for their future. All members of the English department are required to take 4 semesters of ER classes to graduate. The English department students at this institution are assigned to classes mainly based on their yearly TOIEC test

scores, and placed into 2 different ER classes (a higher level *a* class and a lower level *b* class). The average TOEIC score of the *a* class in this case was 560.29 while the average score of the *b* class was 409.85. 12 participants came from the *a* class and 13 came from the *b* class. As teacher-researcher, the *a* class was taught by me. The *b* class was taught by a colleague. The general curriculum of these 2 classes in terms of reading goals was the same, but supplementary activities varied between the classes as described in Section 3.4 below. As these classes occurred during the COVID-19 pandemic, all communication and data collection was done online through learner management systems, video conferencing applications, or e-mail.

3.1.2 Volunteer Recruitment

Both the *a* class and *b* class were given a short presentation on the purpose of this dissertation's research, and on the expectations of participants. There was no penalty for students who did not choose to participate. Initially, 49 students joined the research in the first engagement but only 25 of the collected data sets were considered viable due to several members dropping the course, failing to complete all aspects of the study, or reporting unreliable data as described in Section 3.3 below.

3.2 Design

The research questions of this dissertation are as follows:

1. What differences in student vocabulary knowledge change are there between an extensive reading course with and another similar course without supplementary vocabulary lessons?

2. How can extensive reading instructors set up their courses and materials to optimize vocabulary knowledge growth?

3. What are the advantages and disadvantages of the SRT as a research instrument to measure vocabulary knowledge?

To answer the research questions, an experiment was constructed to measure and compare the vocabulary knowledge growth of the participants. First, 2 groups were made from the participant pool. The *a* class volunteers (n=12) would undertake ER as usual for 1 semester, but they would also receive explicit supplementary vocabulary lessons each week. They will hereby be referred to as the ER+ group. The *b* class volunteers (n=13) would undertake ER as usual but would receive no supplementary vocabulary lessons. They will hereby be referred to as the ER- group. To measure growth, both groups were asked to complete a self-reporting questionnaire in the first and last class of the semester, separated by 15 weeks. This method of group selection was mainly due to the fact that the primary researcher had much more control of the *a* class, making communication about and confirmation of participation in the weekly supplementary lessons

easier. The ER- group was only asked to complete the 2 questionnaires, and the management of their ER throughout the semester was left to their instructor. This experimental set-up is illustrated in Table 3-1 below.

Experimental Element	ER+ Group	ER- Group
Standard ER curriculum	+	+
Supplementary Vocabulary Lessons	+	-
Pre- and Post-semester Questionnaires	+	+

Table 3-1. The experimental set-up.

3.3 Questionnaire Construction and Procedure

To answer Research Question 1, it was known that vocabulary knowledge was the key element that had to be measured, but there appeared to be several theoretical scales upon which to do this, as mentioned in Section 2.1.4. The scales with the most background literature, and the most used with SRTs, were those of receptive and productive vocabulary knowledge (Waring, 1999). Therefore, it was decided that these scales would be separated and used not only to compare vocabulary knowledge growth between classes, but also to compare the 2 types of theorized vocabulary knowledge within individuals or groups. The questionnaire that was constructed for this dissertation borrowed heavily from Waring's (1999) RPV SRT shown in part below in Table 3-2.

Word	Understanding				Use			
accord	0	1	2	3	0	1	2	3
adore	0	1	2	3	0	1	2	3
apply	0	1	2	3	0	1	2	3

Table 3-2. An RPV state rating task (Waring, 1999: 102). Participants were given the an explanation that the numbers in the cells are representative of the following states: 0=unable, 1=can quite well, 2=can well, 3=can very well.

While Waring (1999) had his participants complete this task on paper, due to the online context of the classes in the current investigation, paper questionnaires were impractical. For this reason, the online survey service Google Forms was used to construct an analogous instrument. The questions were also translated into the participants' first language, Japanese, to maximize understanding. Participants were told verbally, through a Microsoft Powerpoint slide, and through instructions on the questionnaire to do the questionnaire alone and without the use of aids. They were also told the meanings of the scales and rating levels verbally in class and in writing at the beginning of the questionnaire. Participants were informed that the receptive scale represented their ability to comprehend the word in English listening or reading, and the productive scale represented their ability to use the word in English speaking or writing. They were asked to select the rating that best described their current state. An example question can be seen below in Figure 3-1.

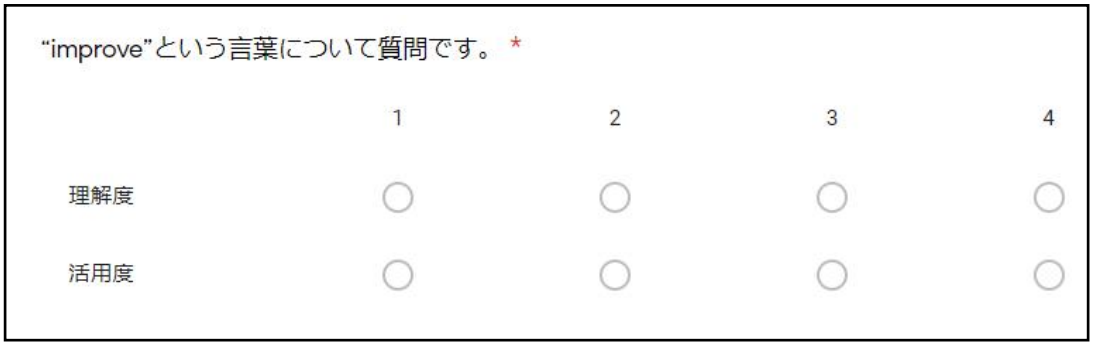

Figure 3-1. An example question from the SRT questionnaire. The text can be translated as: "This question is about the word 'improve'". The scales can be translated as comprehension (top) and utilization (bottom). Regarding the scale, instructions were given in Japanese to participants which were similar to Waring's SRT in Table 3-2: 1=I am unable, 2=I am slightly able, 3=I am able, 4=I am very able.

As one of the advantages of the SRT instrument is that it is quick and efficient, many words can be tested in a short timeframe (Waring, 1999). Therefore, 99 words were included. To compare with similar studies that used the Vocabulary Levels Test (VLT) (such as Webb & Chang, 2015b), the words used in the questionnaires of this dissertation were selected from the same source; Paul Nation's (2017) combined British National Corpus (BNC) and Corpus of Contemporary American English (COCA) list, which are together known as the BNC-COCA list.

The BNC-COCA list contains much more than 99 words, and so a selection process was constructed. The VLT contains a ratio of 3:2:1 for nouns, verbs, and adjectives, and therefore the current investigation's questionnaires also used similar ratios. If a word could be considered more than one part of speech, then the part which it appeared as in the ER class library was its main part. Also, since this dissertation aims to observe change, there is the potential problem that words

that are too difficult or too well known are unlikely to improve. Therefore, through observing the graded readers in the ER library and the past work of students in the program, it was determined that words in the second frequency band (the 1001-2000 most frequently used words) would be an ideal focus for participants in the current study to have partial, but not complete knowledge of the words at the beginning of the research period. Further, as the ER- group were likely to be exposed to words only through reading the graded readers, which were restricted to the available library, potential words were cross-referenced with books in the library to ensure they appeared. If a word appeared at least once in the most popular books read by the previous year's students, then it was considered a candidate for the questionnaire. The most popular books were taken from a list of books completed by 20 or more students in the past year.

One of the issues that informed the vocabulary used in the questionnaire is that many Japanese university students already have perceived knowledge of many words, such as Japanese loan words from English (e.g., *butter* or *model*) or words that students frequently use in junior high school and high school (e.g., *hobby* or the months of the year). Including such words in the research instrument for this dissertation would likely result in high levels of knowledge in the initial questionnaire, which makes showing improvement or growth difficult and therefore they were avoided. If a word was determined to be of this type, it was excluded from consideration, with exceptions made for words with particularly difficult spelling or spelling that does not match the Japanese loan word pronunciation, such as *blonde* or *theater*. The completed list of 89 core words is reproduced in Appendix I. Finally, 5 easy words and 5 pesudo-words were added

as controls that were expected to remain in the high knowledge and low knowledge states respectively. The easy controls were selected as some of the easiest and most frequent words from the 0-1000 frequency BNC-COCA list, while the pseudo-word controls were taken from an online pseudo-word list (Ponting, 2021), both in a 2:2:1 noun to verb to adjective ratio. It was decided that if a participant had a combined number of 5 or more *3* and *4* ratings for pseudo-word controls out of a maximum of 10 (5 receptive and 5 productive controls) in any single questionnaire, then their data would be considered unreliable and inadmissible for use in this dissertation's investigation. Similarly, a combined number of 5 or more *1* and *2* ratings for the easy word controls in any single questionnaire would render a participant unreliable. Finally, the pre- and post-questionnaires had the exact same content of these 99 words, but the order of the words was randomized each time.

3.4 Extensive Reading Class Conditions

The conditions of the ER classes are explained here. Students in the ER classes read e-books via Xreading (URL in the references section) due to the ongoing COVID-19 pandemic which limited their use of the physical library and paper books. Xreading provides students with a library of over 1300 e-books that they can read at their leisure from a wide variety of publishers. Upon completion of a book, students do a short quiz to validate sufficient comprehension of the book, with the quiz pass rate set to 60%.

As part of their course students were required to read a minimum number of words each week of the semester, with their grades penalized by a percentage if

they failed to meet the weekly minimum. This minimum began at 7000 and increased by 300 words each week. A percentage of their grade was also based on the total number of words they read by the end of the semester, with the maximum awarded at 300,000 words. The curriculum of the ER classes observed in this dissertation is the third of a 4-part series that has seen many students achieve the maximum word goals over the past several years. These goals were the same for the ER- and ER+ classes, but supplementary activities were different among groups. The ER- class had several book discussions among classmates while each student in the ER+ class did 3 written book reports individually.

3.5 Supplementary Vocabulary Lessons

The supplementary vocabulary lessons for the ER+ group took the form of short videos, as the class itself was executed through on-demand online videos due to the ongoing COVID-19 pandemic. In each video, students were introduced to 4-8 words from the questionnaire list described in Section 3.3 through Microsoft Powerpoint slides and commentary. The lessons attempted to encompass all aspects of the word, including spelling, pronunciation, syllables, multiple meanings, common collocations, variations in the words' forms, visual aids, and suitable Japanese translations. An example slide can be seen in Figure 3-2 below.

Figure 3-2. An example slide for an ER+ supplementary vocabulary lesson.

It was hoped that this broad-spectrum of instruction would provide a good basis for students to rate an increase in their receptive and productive vocabulary knowledge during the second questionnaire. To further support the productive knowledge growth and to confirm that students in the ER+ group had watched the video, each week they were asked to create several original sentences using the new words as homework. In total, the ER+ group watched 12 videos of about 15 minute length.

3.6 Data Analysis

Data analysis of the SRT questionnaire data first involved identifying overall class patterns in state changes from the pre-questionnaire to the post-questionnaire by constructing a transitional T1-T2 matrix as discussed in Section 2.1.5. This was done for both the receptive and productive knowledge parts of the questionnaires. An aggregate matrix was constructed for the ER- and

ER+ groups by adding all individual student results together. The 5 pseudo-word controls were not included in the SRT matrices or further analysis, as they did not appear in any books or explicit vocabulary lessons. They were analyzed separately for reliability checks only. The easy control words were included in the data analysis as they were expected to occur frequently in the graded readers of the ER classes. This brought the total number of words in each matrix to 94.

Past the analysis of the raw matrix data, 2 statistical methods were used to answer the first research question concerning differences between the groups. These methods aimed to measure vocabulary knowledge growth of the groups and compare them. The data acquired in this investigation can be considered ordinal data by most definitions, for which many researchers caution against using standard parametric data analysis techniques such as means and t-tests (Waring, 1999; Townsend & Ashby, 1984; Jamieson, 2004), and recommend using statistics like median and rank instead. Conversely, there are researchers of the opinion that if an assumption of equal-interval distance between points on the measurement scale can be made, then the mean can have statistical value and may be used with robust statistical tests as an approximation (Sarle, 1995). The questionnaire used in the current investigation had 4 levels with descriptors in the Japanese language that when translated into English match the meanings of: *unable*, *slightly able*, *able*, and *very able*. This dissertation proceeds with the assumption that these points exist on a linear interval scale with approximately equal distance between adjacent states.

Because of this assumption, the SRT data was analyzed through 2 methods of statistical analysis which approach the data from differing angles to check for agreement between their conclusions. These methods were an analysis of the variance of group mean ratings, often used for data on a linear scale, and an analysis potential growth achieved, a relatively new form of analysis for ordinal data. A similar combination of methods was used by Zhang et al. (2021) to measure subjects' abilities and self-confidence in using an instrument in the field of ophthalmology. The same rationale should be applicable to measuring vocabulary knowledge, but it should be made clear that Zhang et al. (2021) used an actual test of skill with a linear scale in their analysis of variance (ANOVA) analysis, and measured potential growth achieved in self-confidence with a separate self-reported ordinal scale. The current investigation attempts to analyze both on the SRT scale alone. Further true tests of skill were not as feasible in this case due to the global COVID-19 pandemic, which is discussed as a limitation in Chapter 5.

3.6.1 ANCOVA Investigation of Differences in Post-Questionnaire Ratings

First, using the mean rating for each individual, a one-way analysis of covariance (ANCOVA) was performed in IBM's SPSS software to investigate the post-questionnaire ratings for significant differences between the ER- and ER+ groups for both RPV knowledge types. The pre-questionnaire results were used as controlling covariants. This method was selected as it can measure differences in the post-questionnaire ratings between groups while controlling for the differences in initial ratings among subjects in the pre-questionnaire—a property often desired for pre/post-test investigations where the participants in the

groups are not randomly selected (Woodrow, 2014). ANCOVA does not investigate differences between the pre-questionnaire and post-questionnaire directly, nor does it check for differences between groups in pre-questionnaire ratings; however, the prerequisite checks for the ANCOVA test do provide some insights into some of these areas. The first check is that there is no statistically significant difference between groups on the pre-questionnaire, which was investigated through a regular 1-way analysis of variance (ANOVA) in SPSS. The second is the check for homogeneity of regression which was done by showing that there was no statistically significant interaction between the group variable and the pre-questionnaire. This was investigated through an ANCOVA interactions model in SPSS. The third check is for homogeneity of variance, which was checked with Levene's Test of Equality of Error Variances in SPSS. The usual parametric test requirement of normality in the dependent variable, mean SRT rating in this case, remained as an assumption due to the low group sample sizes of 12 and 13 and is one of the limitations of this investigation, although Shapiro-Wilk tests were used to provide some evidence for normality.

3.6.2 Investigation of Potential Growth Achieved

The second method by which the data was analyzed involved considering the potential increases and decreases that could have occurred within subjects. A method of calculating indicators for potential positive change achieved and potential negative change achieved for items in ordinal data sets has been described by Ferreira et al. (2013). These indicators are called the indicator for positive change (I_PC) and the indicator for negative change (I_NC). The items analyzed were people in the original case, but the items are words in the case of

this dissertation. This method involves calculating a ratio of positive or negative changes over the total number of states which could have changed positively or negatively at T1. This method also weights particular changes more than others. In the base formula, which is used in this dissertation, the weights assume equal distance between adjacent the states. This method still relies on the assumptions made in Section 3.6 for accuracy, yet it it does not require the assumption of a complete continuous linear-interval scale which averaging requires and may therefore be a better approximation. Ferreira et al.'s (2013) general equations for I_PC and I_NC are shown in Figure 3-3 below.

$$I_PC = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} C_{ij}(j-i)}{\sum_{i=1}^{m-1} N_i(m-i)}$$

$$I_NC = \frac{\sum_{i=2}^{m} \sum_{j=1}^{i-1} C_{ij}(i-j)}{\sum_{j=1}^{m-1} N_j(m-j)}$$

Figure 3-3. The general case equations for I_PC and I_NC, where m is the number of states on the scale, i and j are the numbers of the rows and columns of the matrix containing the number of items that changed state (e.g. Table 2-2), C_{ij} is the number of items in a particular cell of the ij matrix, N_i is the total number of items in row i, and N_j is the total number of items in column j. The denominator has been adjusted to account for the fact that the rows in this dissertation's matrices represent T1, while in the original Ferreira et al. (2013) publication they represent T2.

These calculations can be applied to the participants' rating changes; however, a more useful indicator would be one which combines both aspects of positive and negative potential achievement. I propose a new equation for such a value which I call I_TC, the indicator for total change. This equation can be seen in Figure 3-4 below.

$$I_TC = \frac{(\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(C_{ij}-C_{ji})(j-i))((\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}C_{ij})+(\sum_{i=2}^{m}\sum_{j=1}^{i-1}C_{ij}))}{(\sum_{i=1}^{m-1}N_i(m-i))(\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}C_{ij})+(\sum_{i=2}^{m}N_i(i-1))(\sum_{i=2}^{m}\sum_{j=1}^{i-1}C_{ij})}$$

Figure 3-4. A proposed general case equation for I_TC, an indicator of total net change. The variables have the same representations as in Figure 3-3. In this dissertation, m=4.

The properties of the I_TC equation are as follows:

(1) The range of the I_TC equation is from -1 to 1.

(2) If $I_PC=I_NC$, then $I_TC=0$.

(3) If all items are in the *j=m* column, then maximum positive potential has been achieved and $I_TC=1$.

(4) If all items are in the *j*=1 column, then maximum negative potential has been achieved and $I_TC=-1$.

(5) If there are no items, or items fall only on the diagonal *i=j*, then I_TC is undefined.

(6) Other placements of items in the matrix are reflected in the I_TC value with weights of potential achievement attained in both directions being considered.

The method of calculating I_PC, I_NC, and I_TC values above gives some insight into the growth of individuals and classes, but to compare the ER- group with the ER+ group, a Mann-Whitney U unpaired non-parametric test was used to analyze the I_TC values by group. This test was selected as it functions even in the case where the sample sizes are low and the I_TC values cannot be assumed to be normally distributed (MacFarland & Yates, 2016). This provided insight into whether or not one group rated their improvement as significantly higher than the other, in terms of potential growth achieved.

3.7 Summary

This chapter has laid out a brief overview of the participants involved, and all of the aspects of the experiment and data analysis in this dissertation. The ER- group contained lower level Japanese university students who undertook regular ER classes while the ER+ group contained higher level students who undertook ER classes with weekly supplementary vocabulary lessons. To collect data, a 4-point SRT separated into receptive and productive components was used at the beginning and end of the school semester. To analyze the data, the SRT ratings were assumed to be on a linear scale and the post-questionnaire mean ratings were compared through traditional parametric statistical methods. Potential growth achieved was also investigated through non-parametric methods designed for ordinal data to check for agreement.

CHAPTER 4

RESULTS OF THE SRT QUESTIONNAIRES AND STATISTICAL TESTS

In this chapter, the results of all collected SRT data are displayed in matrix form. Following this, the results of several statistical tests that were performed are presented and interpreted.

4.1 Unsuitable Data

All data was collected from the pre- and post-questionnaires, and was first assessed to determine if a given subject was reliable. As stated in Section 3.3, if a participant had 5 or more combined *3* or *4* ratings for the pseudo-word controls, or *1* or *2* ratings for the easy word controls, in any single questionnaire then they were disqualified from further analysis on the grounds of reliability. This resulted in 8 disqualifications from the ER- class, and 4 from the ER+ class, bringing the group numbers to n=13 for ER- and n=12 for ER+ respectively. Along with subjects who abandoned the ER courses or research group, this significantly lowered the number of participants from the original group of 49 who took the pre-questionnaire. This process was not ideal from a sample size perspective, but the remaining data had a much greater degree of reliability.

4.2 SRT Matrices

The data is presented as state changes from pre-questionnaire (T1) to post-questionnaire (T2) in matrix format, as was shown and described in Table

2-2 of Section 2.1.5. Matrices can be constructed for each participant and summed together to get an overall view of each group. These group matrices for all groups and types of vocabulary knowledge can be seen below in Table 4-1. Group data is presented in this chapter, but 2 single participant examples can be seen in Appendix II.

ER- (R)	(→)State I	(→)State II	(→)State III	(→)State IV	ER- (P)	(→)State I	(→)State II	(→)State III	(→)State IV
State I	25	16	29	29	State I	59	97	43	20
State II	10	46	76	44	State II	30	167	74	45
State III	12	40	148	137	State III	17	109	94	65
State IV	6	16	88	500	State IV	10	58	103	231
ER+ (R)	(→)State I	(→)State II	(→)State III	(→)State IV	ER+ (P)	(→)State I	(→)State II	(→)State III	(→)State IV
State I	14	13	13	25	State I	27	58	42	27
State II	4	17	24	44	State II	5	91	86	87
State III	3	14	70	232	State III	6	55	127	182
State IV	1	6	70	578	State IV	0	18	69	248

Table 4-1. Summed T1->T2 state changes in receptive (R) and productive (P) vocabulary knowledge ratings for the ER- (n=13) and ER+ (n=12) group. Row labels represent the states at T1, and column labels represent the states at T2. The top-left to bottom-right diagonals represent states which did not change, coloured in gray. Cells to the left of the diagonal represent decreases in state, coloured in orange, and cells to the right represent increases in state, coloured in green (N=25 students, 94 words each).

The results of these matrices appear as expected with generally net positive growth and high concentrations near the diagonal. The comparatively high numbers of State I(→)State IV for receptive knowledge in both groups may be one point of interest. The ER- group also reported much greater numbers of decreases in state than was expected for productive knowledge. In both groups, the high numbers of State IV(→)State IV static results for receptive knowledge suggests that a more difficult selection of words likely would have been more appropriate.

4.3 Growth Comparisons Between Groups

The first research question asks what differences in vocabulary knowledge change there are between the two groups. This section presents the 2 statistical methods that were used to investigate differences between groups. The descriptive statistics of the SRT ratings are shown followed by the ANCOVA analysis of variance tests. Then, the calculated indicators for positive, negative, and total net change are shown before presenting the results of the Mann-Whitney U test comparisons.

.

4.3.1 Descriptive Statistics and ANCOVA

First, the basic descriptive statistics for the RPV knowledge ratings of all individuals were calculated in Microsoft Excel. The full results are shown in Appendix III. The overall mean ratings of the groups are then shown in Tables 4-2 and 4-3 below.

Group	Mean	Std. Deviation	*n*
Er- (PrQ)	3.19	0.36	13
Er- (PoQ)	3.40	0.36	13
Er+ (PrQ)	3.39	0.37	12
Er+ (PoQ)	3.70	0.3	12

Table 4-2. The group means and standard deviations of the individual mean ratings for receptive vocabulary knowledge responses (N=25). PrQ=Pre-questionnaire, PoQ=Post-questionnaire

Group	Mean	Std. Deviation	*n*
Er- (PrQ)	2.71	0.42	13
Er- (PoQ)	2.75	0.55	13
Er+ (PrQ)	2.78	0.39	12
Er+ (PoQ)	3.21	0.47	12

Table 4-3. The group means and standard deviations of the individual mean ratings for productive vocabulary knowledge responses (N=25). PrQ=Pre-questionnaire, PoQ=Post-questionnaire

Using the mean rating for each individual, a one-way analysis of covariance (ANCOVA) was then performed in SPSS to investigate the post-questionnaire ratings for significant differences between the ER- and ER+ groups for both RPV knowledge types. The pre-questionnaire results were used as controlling covariants. The main assumptions of this test were met, in that there were no statistically significant differences in the pre-questionnaire group ratings one-way ANOVA (receptive p=0.204, productive p=0.725), there were no significant interactions between the group and pre-questionnaire in the interaction model (receptive p=0.483, productive p=0.710), and Levene's Test showed no statistically significant deviations from the null hypothesis that the error variance of the dependent variable was equal across all groups (receptive p=0.413 , productive p=0.624). The assumption of normality in the data remains an approximation due to the low sample sizes of n=12 and n=13 in the groups, although Shapiro-Wilk tests showed no statistically significant deviations from normality (p values for all groups and knowledge types ≥0.05).

The ANCOVA tests revealed a statistically significant difference between the ER- and ER+ groups ($F(1)=10.959$, $p=0.003$), with a large effect size of $\eta_p^2=0.332$, according to Cohen's (1998) classification of effect sizes, for productive vocabulary knowledge ratings. The difference for receptive vocabulary knowledge ratings was not statistically significant, $p=$ (0.097). The tests of between-subjects effects are reproduced in Appendix IV. These results can be summarized by concluding that although both groups had no statistically significantly difference among pre-questionnaire ratings, the ER+ group rated themselves significantly higher in productive vocabulary knowledge in the post-questionnaire than the ER- group when controlling for pre-questionnaire answers. No such difference was found in the receptive knowledge ratings.

4.3.2 Analysis of Potential Growth Achievements

The second method by which the SRT data was analyzed involved considering the potential increases and decreases that could have occurred within subjects. The I_PC, I_NC, and I_TC values for receptive and productive vocabulary knowledge ratings for each student were calculated in Microsoft Excel and compiled in Tables 4-4 and 4-5 respectively.

Participant	I_PC	I_NC	I_TC
ER-			
1	0.8971	0.3000	0.7829
2	0.1818	0.3636	-0.1261
3	0.3289	0.3108	0.0142
4	0.7245	0.1000	0.6669
5	0.3833	0.2600	0.0853
6	0.8400	0.3333	0.3942
7	0.3894	0.1977	0.2075
8	0.2385	0.2385	0.0000
9	0.3333	0.3506	-0.0138
10	0.6364	0.3846	0.0888
11	0.5185	0.5806	-0.0505
12	0.4412	0.2963	0.1122
13	0.5674	0.0758	0.5319
ER+			
14	0.8182	0.4286	0.2003
15	0.5455	0.2857	0.1336
16	0.9600	0.0000	0.9600
17	0.7632	0.1818	0.4686
18	0.9107	0.0000	0.9107
19	0.5243	0.2222	0.3384
20	0.6667	0.1000	0.6050
21	0.4667	0.8000	-0.1127
22	0.8852	0.3000	0.7322
23	0.3782	0.1868	0.2121
24	0.5135	0.6087	-0.0523
25	0.8933	0.0000	0.8933

Table 4-4. The receptive I_PC, I_NC, and I_TC values for each individual (N=25).

Participant	I_PC	I_NC	I_TC
ER-			
1	0.6346	0.1364	0.5863
2	0.2449	0.2745	-0.0265
3	0.1879	0.2143	-0.0356
4	0.3909	0.0219	0.3887
5	0.1724	0.3628	-0.1584
6	0.3000	0.7328	-0.4017
7	0.1152	0.1512	-0.0499
8	0.2707	0.1565	0.1306
9	0.3588	0.2759	0.1073
10	0.4333	0.3704	0.0213
11	0.3919	0.5755	-0.1907
12	0.2742	0.2500	0.0280
13	0.1720	0.2121	-0.0576
ER+			
14	0.7778	0.2308	0.4336
15	0.4342	0.2951	0.1184
16	0.8780	0.1667	0.8072
17	0.2655	0.2455	0.0214
18	0.5208	0.1266	0.4540
19	0.5259	0.0986	0.4705
20	0.5748	0.1940	0.4540
21	0.4925	0.3704	0.1071
22	0.5813	0.1067	0.5425
23	0.3578	0.0438	0.3476
24	0.3636	0.3636	0.0000
25	0.3475	0.2596	0.1028

Table 4-5. The productive I_PC, I_NC, and I_TC values for each individual (N=25).

Using the I_TC data from Tables 4-4 and 4-5, Mann-Whitney U tests were performed in SPSS to determine whether or not there was a statistically significant difference between the ER- and ER+ groups with regards to potential growth achieved. A graphical representation of frequencies can be seen in Figures 4-1 and 4-2, and summaries are presented below. The full results of the tests can be seen in Appendix V.

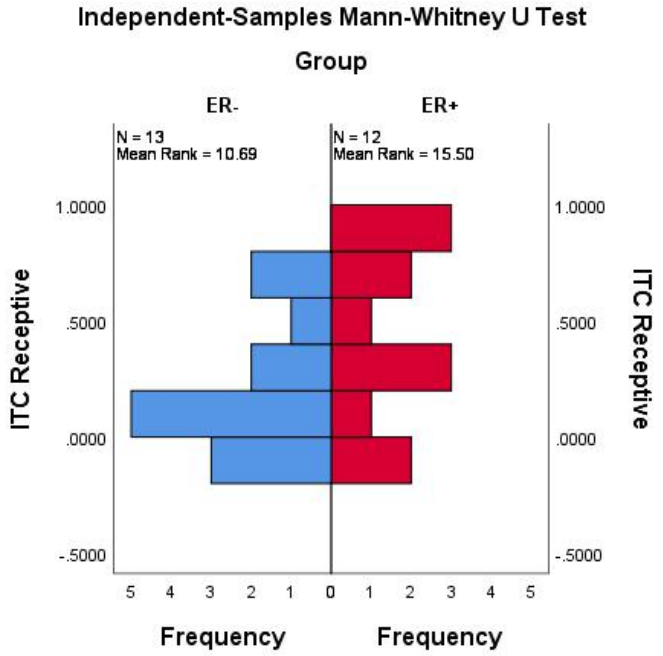

Figure 4-1. A graphical representation of the frequency of receptive I_TC values for the ER- and ER+ groups (N=25).

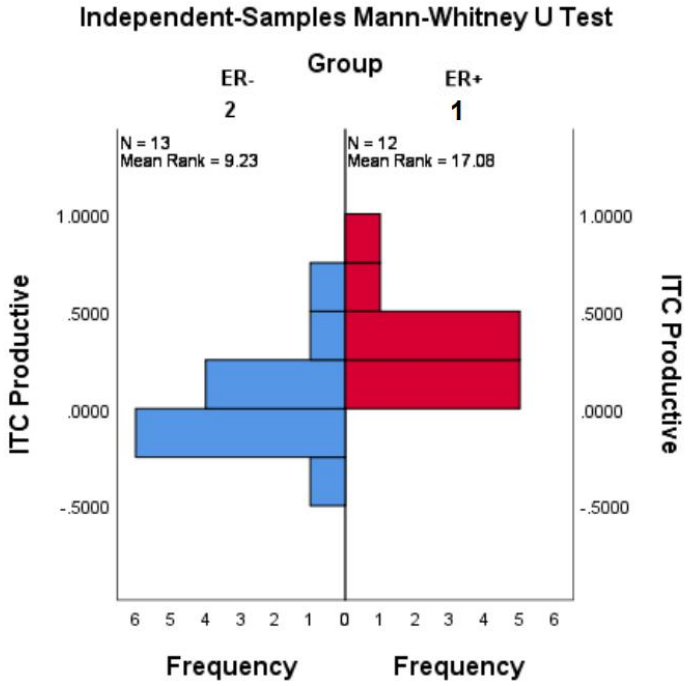

Figure 4-2. A graphical representation of the frequency of productive I_TC values for the ER- and ER+ groups (N=25).

The results of the Mann-Whitney U tests indicate that for productive vocabulary knowledge, the ER+ group had more positive potential growth achieved than the ER- group to a statistically significant degree, p=0.007, with a Pearson's correlation effect size of r=0.533, which is a large effect according to Cohen's (1992) classification of effect sizes. Conversely, there was no statistically significant difference for receptive vocabulary knowledge ratings, p=0.110. This is in agreement with the mean ratings ANCOVA analysis in the previous section and further suggests that this result closely approximates the reality of the situation if prior assumptions are accurate.

4.4 Summary and Significance of the Findings

This section concisely discusses the findings of Chapter 4, and to what degree they should affect the conclusions in the following chapters. First, the raw data obtained from the questionnaires and the SRT matrices that they generated were discussed. The unsuitability of a large percentage of the raw data was addressed, and with this data removed the remaining data should include ratings that more accurately reflect the true knowledge states of the participants. The SRT matrices are consistent with the expectations that most shifts would be concentrated around the top-left to bottom-right diagonal (Waring, 1999), indicating mostly static ratings or near state changes. Generally, more shifts appeared to the right of this diagonal line, which indicates greater perceived vocabulary knowledge at the end of the semester. This would be expected for university students who are learning English in a variety of classes and participating in ER or vocabulary lessons, although productive ratings of the ER-

group also had a high number of negative shifts which may be a consequence of insufficient input or practice with usage.

The first research question concerns the differences in change between the 2 groups due to treatment. Findings of a statistically significant difference would indicate that something in the difference in treatment for these 2 groups may have brought about this difference. The ANCOVA test procedures did reveal such a difference, but they proceeded with an assumption of a normal distribution in the data that is unlikely to be met by this small sample size, although analysis of variance tests are generally robust towards normality violations (Lix et al., 1996). On the other hand, the Mann-Whitney U non-parametric comparison test used in Section 4.3.2 does not rely on the assumption of a normal distribution, and should be capable of confidently detecting a difference in the groups if the effect size is large enough. Both statistical methods were consistent with each other in concluding that the ER+ group showed a degree of higher positive growth than the ER- group to a statistically significant degree with large effect sizes when it came to productive knowledge ratings, but not for receptive knowledge ratings.

In summary, if the assumptions of equal distance between points on the SRT rating scale and participant ability to self-report accurately hold true, then a group with supplementary vocabulary lessons will likely have higher productive vocabulary knowledge gains over a similar time period to that observed in this dissertation. The same cannot be said for receptive vocabulary knowledge at this time. Possible reasons for these results are discussed in the following chapter.

CHAPTER 5

DISCUSSION REGARDING VOCABULARY IN EDUCATION AND THE SRT IN RESEARCH

This chapter aims to discuss what the answers to the research questions mean for education moving forward, the limitations of the study, and the advantages and disadvantages of the SRT instrument for measuring vocabulary knowledge as demonstrated in this dissertation.

5.1 Insights on Vocabulary Knowledge Measurement and Learning Input

Many of the findings in this dissertation are not new discoveries, but they do align with other previous findings in the literature and widely held beliefs. Two of such beliefs are the idea that receptive knowledge is greater than productive knowledge (Myers, 1914; Melka, 1997; Nation, 2001), and the idea that intentional teaching or a mixed style of input is more efficient than strictly incidental doses of input (Meganathan et al., 2019; Waring & Nation, 2004). The the raw SRT ratings solidly back up the former claim, as receptive knowledge ratings were almost always higher than productive for individual words and the descriptive statistics of the SRT ratings showed showed the dominance of receptive knowledge ratings for any given time or group. Further, the results of the ANCOVA and Mann-Whitney U tests for group differences fairly strongly back up the latter claim as the ER+ group reported higher post-questionnaire ratings and higher potential growth achieved in the positive direction for productive knowledge than the ER- group to a statistically significant degree.

The idea of the I_TC for use in data analysis is a relatively new one that I have not encountered in the literature for self-reporting vocabulary knowledge, but it seems that the I_TC adequately functions as a value that can be used in statistical analysis that encompasses a subject's tendency to gain or lose vocabulary knowledge. The I_PC and I_NC indicators have been used in medical fields (Ferreira et al, 2013; Zhang et al, 2021), but in the case of studying vocabulary knowledge in language education I believe a net total indicator is more useful than either of the individual 1-way indicators because considering one without the other can be misleading. The validity of the self-reported I_TC variable should be verified by comparing with other accepted testing instruments to further confirm that it actually represents a measure of vocabulary knowledge growth, and if it is deemed valid then I believe it could be a very useful value for other uses with the SRT instrument, along with strengthening the conclusions of this dissertation.

SRTs have previously been analyzed by separating changes into near or dramatic state changes depending on whether they were a shift of 1 state or greater than 1 state (Waring, 1999; Dabaghi & Rafiee, 2012). This is essentially taking a step from nominal data to ordinal data by suggesting that some states are closer or further than others with regards to a decided starting point. The I_TC calculations derived from Ferreira et al.'s (2013) positive and negative indicators takes this a step further by weighting each step of the scale. This approach requires a bigger assumption about the distances between the states, which is that the distance between adjacent states is equal, but offers more detailed evaluations if the assumption is accurate. If the scales used in future investigations can be argued to have equal distances between adjacent states, I believe that the I_TC is

better metric to measure change than counting near and dramatic state changes. Further, the I_TC equation can be modified if states are thought to have varying distances. For example, instead of weighting the step between *I am unable* and *I am slightly able* as 1 (signified by *j-i*, or 2-1 in this case), this could be modified to 1.5 or 2 if the researcher wishes to weight this step as a greater distance covered. This gives flexibility to the I_TC approach and allows for improved calculations if more is known about the states being used.

Several other points of interest in the results should be considered. First, In both the ANCOVA and Mann-Whitney U tests, a significant difference was detected between the groups for productive vocabulary knowledge ratings but not receptive. There could be several reasons for this, and one of them may be that the only tasks that any of the participants were asked to do outside of the questionnaire was when the ER+ group was asked to produce sentences with the new words they observed every week as a completion check. Second, both groups had exceedingly high numbers of *4* ratings when rating receptive feedback at T1, indicated by the high averages of 3.19 (ER-) and 3.39 (ER+). This means that there was less room on the scale to show improvement in receptive vocabulary knowledge. This could be a reason why receptive knowledge ratings did not appear to grow as much, or a reason why they did not appear significantly different between groups. The word selection in the questionnaires for these groups appeared adequate for measuring productive knowledge changes, but it was likely too easy to adequately measure receptive knowledge changes. This should serve as a caution to researchers that the same set of words may not always be adequate to measure both of these knowledge types. In this case, it

seems that the differences between the ER- and ER+ groups were more pronounced when looking at reports of productive knowledge, and this could be an indicator that researchers and educators should prioritize measuring productive knowledge when they desire higher sensitivity in observing changes; however, a more difficult set of words may reduce these differences.

The 2 most relevant differences between the ER- and ER+ groups were the explicit vocabulary lessons that were provided to the ER+ group, and the higher average TOIEC English test scores of the ER+ group, indicating a higher level of English proficiency at T1. It is likely that the reason for the differences in mean ratings and potential growth achieved between groups is most likely a combination of both factors, as explicit vocabulary lessons could clearly be a major factor in increasing the knowledge of words, but also because a high score on a language test like TOEIC is likely to correlate with other characteristics that are desirable for learning such as a good memory (Bosman & Janssen, 2017), high motivation to continue studying English (Suzuki & Sogawa, 2010), or other personality traits such as orderliness and self-confidence. The lack of significant differences in the pre-questionnaires between groups helps to strengthen the idea that the main contributor to later differences was the vocabulary lesson treatment, but the experimental set up of this dissertation means that it is not possible to definitively determine which of these factors was the main contributor to differences in change. It can be said that a group with all of the advantages of the ER+ group is likely to have higher self-reported growth rates for productive vocabulary knowledge than a group which does not.

5.2 Charting a Way Forward for Vocabulary Input in Education

The second research question asks how instructors can set up their courses and materials to optimize vocabulary knowledge growth in education. Given the insights and conclusions discussed in the previous section, I will recommend a way forward in this regard.

The findings of this dissertation support the ideas of Waring & Nation (2004) and Meganathan et al. (2019) in that balanced sources of input are more efficient than input mainly coming from incidental sources. From the agreement with this conclusion, I then propose that students be tested early on with an instrument based on the BNC-COCA frequency list to determine the level of their vocabulary knowledge. This may be a Vocabulary Levels Test (McLean & Kramer, 2015; Schmitt et al, 2001), or some kind of self-reporting instrument such as the one used in this dissertation. Then, words and word families from frequency lists corresponding to the students' levels of vocabulary knowledge should be explicitly and regularly incorporated into their reading classes or other areas of their required curriculum. The optimal nature of this incorporation is not yet known, and although this dissertation has shown that a mostly passive approach of watching videos may be enough to produce a significant difference on its own, incorporating productive activities and practice is likely to offer further significant advantages if time allows.

Furthermore, as many graded readers that I have encountered during the execution of this research do, I highly recommend that authors and publishing companies include supplementary vocabulary definitions, practice activities, and

visual aids for words that occur that are a step above the general word frequency level of the graded readers which contain them. I also encourage instructors to use these resources, or modify them to their needs, as it would be an easy way to move towards a better balance of incidental and intentional input. The Xreading ER website has begun to add pre-reading vocabulary assignments for each graded reader that instructors can use alongside traditional reading and quiz assignments, which is a step in the right direction. Instructors can also incorporate similar activities for paper books using written or oral assignments that add explicit focus to difficult or new words.

5.3 Limitations

Although this research produced valuable perspectives and comparisons, there were 5 main limitations that should be discussed. They are the group selection process, the final sample sizes, the assumptions of linearity, the sampling frequency, and the restrictions due to the COVID-19 pandemic. Each of them are described and potential alternative methods or solutions to the problems are suggested.

5.3.1 Group Selection

The first limitation was the group selection method and it relates to why the reason for the significant differences between the 2 groups could not be isolated. This was mainly because the ER- and ER+ groups consisted of lower level students and higher level students respectively, with grouping based mainly on their TOEIC test scores. It is not known whether the main reason for the ER+

experiencing higher growth was due to their intrinsic ability as a higher level group, or whether it was due to the explicit vocabulary lessons. This difference in ability was known in the planning stages; however, any pair of classes taken from this particular university would have some difference in year, academic department, or level. The level difference was deemed to have the least amount of an effect among these differences, which is why classes from the same year and department but differing levels were selected. In retrospect, it may have been better to randomly assign volunteers from both of these classes into the ER- and ER+ groups. This was not done because a different instructor was managing the other class, but this could be overcome with some coordination and it would have provided a more useful conclusion. Another option would be to use the inherent differences between these classes as a kind of initial treatment, and forego the explicit vocabulary lessons. This would provide a clear indication of whether or not the difference in class level is enough to produce a significant difference between groups, but removes the ability to answer questions about incidental versus intentional input.

5.3.2 Sample Size

Several causes led to a final sample size that was smaller than originally expected. Going into the research, it was known that the maximum potential group size for the ER- and ER+ groups was 40 members each. This would have been a very adequate group size, but the process of recruiting volunteers was the first significant cut to the group sizes, bringing them to 26 and 23 respectively. This is much less desirable, but still likely would have been adequate for this research as a sample size of around 30 is often quoted as necessary for

confirming a normal distribution depending on the context (Boos & Hughes-Oliver, 2000; Krithikadatta, 2014). More effort could have been made to ethically encourage students, or other classes could have been invited for recruitment. The second significant reduction in sample size came from participants who did not follow through with completing the second questionnaire. Unsurprisingly, this was a much greater issue with the ER+ class, as they were required to watch videos and submit short assignments weekly which some participants likely found troublesome and there was minimal incentive to continue. Both groups also had students who dropped the course and could no longer participate. The third significant reduction was due to students who failed the reliability control check discussed in Section 4.1. It was initially thought that this check would exclude 5 to 10 percent of participants from providing usable data, but in reality this number was closer to 50 percent of the remaining groups. This may be due to problems with the SRT questionnaire process that are discussed in the following section.

5.3.3 Assumptions of Linearity

The results from the 2 statistical methods used in this dissertation were in agreement, which strengthens the overall conclusion, but the calculations of mean ratings for the ANCOVA analysis along with the calculations of the I_TC values both assume linearity of the self-reporting scale used in the questionnaire to a certain degree. Most of the conclusions of this dissertation rely on these assumptions, but more conservative thinkers of statistical analysis may say these assumptions do not have merit (Waring, 1999; Townsend & Ashby, 1984; Jamieson, 2004). If a statistical method for analyzing statistical differences in

ordinal data exists that does not rely as much on these assumptions, it should likely be used as an alternative to the methods used in this dissertation. One potential example of this is the idea of effect size (ES) and its various versions that are used in health fields of academia for analysis of ordinal data (Middel & Van Sonderen, 2002).

5.3.4 Sampling Frequency

The research of this dissertation only used two data points in time through measurements with the pre-questionnaire and post-questionnaire. Other studies using SRTs (Waring, 1999; Dabaghi & Rafiee, 2012) often used more time points to paint a clearer picture of change and to reduce the effects of single time point inaccuracies. I feel that it would not have been difficult to include 1 or 2 more data collection points throughout the semester which could have strengthened the conclusions of this research.

5.3.5 COVID -19 and the Online Teaching and Learning Environment

While carrying out the research of this dissertation online certainly streamlined many logistical aspects, it severely altered the natural course of interaction with subjects as well as limiting observations and supervision. Outside of pandemic times, I feel it would be beneficial to conduct data collections in person and observe at least some of the reading time of subjects. Digital means are still recommended over pen and paper, but the instructor should have more of a presence in the study than was had in this research. Further reasoning for this is discussed in the following section.

5.4 Assessing the SRT as a Research Instrument

The third research question asks about the advantages and disadvantages of using the SRT as a research instrument to measure vocabulary knowledge. I was pleasantly surprised by many aspects of using the SRT overall, and although some significant flaws appeared throughout the research of this dissertation I believe they could be mostly rectified with a reasonable amount of caution and proactive action on the part of researchers and educators.

First, the ease of use in terms of instrument creation, administration, and analysis was a highly desirable factor. I feel that the number of 99 words could be scaled up to 200, 500, or even 1000 without much of an increase in effort on the part of the researcher. I expect that participants would still be able to complete the questionnaire in under 30 minutes in such cases, which is still a relatively short amount of time. Using Google Forms made sharing to participants' devices extremely simple, and the automatic conversion into a spreadsheet made counting and analysis very efficient. This ease of use was also reflected on the student side, as it seemed they were able to easily understand how to consider a rating and input their answer quickly on a computer or smartphone.

As described in Section 2.1.3, the benefit of clearly defining RPV knowledge to participants and having them consider and self-report it on a scale bypasses several irrelevant skill checks that many other tests aiming to exclusively measure these knowledge types tend to have. This comes with a caveat, as self-reporting what is known or what abilities are held is not going to be as directly truthful as somehow demonstrating it. Often times we think we know something, but we are

incorrect. However, if the self-reporting and actual demonstrating are the same within acceptable margin then self-reporting may be an extremely efficient alternative that could be used in situations in which traditional testing could not. Furthermore, until other commonly used tests are finely tuned to the point where they can test very precise types of knowledge, which may or may not ever be the case, then self-reporting on a gradient will have the ability to detect knowledge and changes that many more traditional types of tests miss due to their inclusion of other skill requirements, according to Waring (1999: 9-21). The inclusion of controls like pseudo-words can help maintain the reliability of the data, and other ways to do this should certainly be explored. One might use several replicate questions, or interviews to accomplish this.

These advantages seem very positive, but some of them also manifest as disadvantages. The ease of answering the questionnaire can be a great feature, but going through many questions can be monotonous and cause the participant to speed up or lose concentration. This could have negative effects on the accuracy of results due to inadequate consideration time or mistaken inputs. On the day of the post-questionnaire I was able to enter the school and observe several students taking the questionnaire on their smartphones, and I believe I witnessed this phenomenon. Furthermore, 2 students were observed doing the questionnaire at the same time together and chatting, despite the instructions indicating it should be done alone. This could clearly cause inaccuracies and violations of the assumption of independence among data sources for statistics. For these reasons, the participants should be observed during data collection whenever possible.

Bringing the participants' attention to focusing on the task and emphasizing the seriousness of the project may be required. For significantly longer questionnaires of this style I would recommend breaking data collection sessions into 2 or 3 instances to allow for breaks.

I believe the reasons discussed above are what led to the large percentage of participants who failed the pseudo-word controls, and consequently had unreliable data overall. It becomes easy to quickly pass through the questionnaires inputting only *3* or *4,* but for usable data it is necessary for all participants to adequately consider all points on the scale. Waring (1999: 116) also made a cautionary observation about his study using pseudo-words when he wrote, "It is possible that the subjects wanted to show development over the 11 months of the study and may have subconsciously rated their knowledge slightly higher at each datatime." This would indeed be a serious problem, and it could explain why more subjects failed the controls at T2 in this dissertation. This effect would most likely be relatively equal between large groups of random samples from the population, and is therefore unlikely to significantly change comparisons between such groups; however, it would have a much larger effect when trying to measure growth within a single group. These issues are mostly related to a multi-level SRT, and while YES/NO checklist SRTs are also possible, they will not give in-depth perspectives into single words which was one of the goals of this dissertation. Consequently, this needs to be considered when planning any given research. Either higher stakes should be placed on taking the SRT questionnaire seriously, or sample size should be large enough that even a large percentage of loss due to failed controls remains adequate. Other methods

of reclaiming this data may also be explored, such as asking students who failed the control to do the questionnaire again with new controls.

Another point worth mentioning is the lack of literature on data analysis methods. The SRT data obtained from in this dissertation could be considered ordinal data similar to a Likert scale, and treating the ordinal data as linear has clear disadvantages. Try as we may to define the rating numbers, each participant constructs his or her own scale in their mind and each one is unlikely to be a linear scale with equidistant points. This means that we may be better off analyzing things like the number of items per state and the changes between states without assigning numerical weights. The best way to statistically analyze this type of data is not clear to me, and although 2 methods assuming linearity or equal distance between adjacant points that were applied in this dissertation appeared to lead to reasonable agreement and success, I believe there is significant room for improvement. Scaling this type of analysis up to a higher number of states or states which are not clearly above or below each other may seem like advantages of using an SRT instrument, but the data analysis will quickly become even more difficult to effectively pilot.

Overall, I am optimistic about the future of the SRT as an instrument and specifically its ability to efficiently probe different types and levels of vocabulary knowledge. Some of the next steps are clear, such as finding the best data collection conditions, while others seem distant, such as engineering how to statistically analyze the ratings to arrive at the most meaningful and accurate conclusions.

CHAPTER 6
CONCLUSION

The results of this dissertation are in agreement with the idea that a mixed method of intentional and incidental learning with explicit and implicit input is better for vocabulary knowledge growth than an implicit input focused method. When it came to productive vocabulary knowledge, one group of higher level extensive reading class students who undertook ER for 1 semester along with weekly vocabulary lessons reported a higher post-questionnaire average and indicator for total change than a lower level group which only undertook extensive reading. This difference cannot be completely contributed to the addition of explicit vocabulary lessons, as the participants were grouped mainly by TOIEC score rather than random selection. Nevertheless, it is recommended that educational content creators and educators continue to incorporate intentional focus points for lesser known vocabulary along with general content for implicit learning, such as extensive reading, until there is conclusive evidence that it is not needed.

The research of Waring (1999) provided us with an eye-opening perspective of the problems of vocabulary knowledge scales, and his proposal of SRTs may be one of the most efficient instruments that we have to measure and observe changes in knowledge over time. Even if SRTs are not widely adapted due to reliability, validity, or data analysis issues, the problems that they attempted to overcome should be key points to consider as the creation of vocabulary knowledge scales inevitably continues.

The immediate next step for research in this area would be to perform a similar experiment on 2 groups of participants selected randomly rather than based on test scores to observe if the same result is achieved. It is also necessary to perform further checks on the validity and linearity of self-reporting scales, including the receptive and productive 4-point scales used in this dissertation. Efforts to check and increase the reliability of subject responses would also be greatly beneficial to the continued use of self-reported data, as a greater proportion of initial participants could be used and researchers could be more confident with the conclusions of their research. Finally, the properties of a book or library that are the most essential to vocabulary knowledge growth should be explored. It would be of interest to see how SRT results correlate with variables such as total input, in terms of words read or time spent reading, or with the difficulty level of input.

References

Bamford, J. and Day, R. 1997. Extensive Reading: What is it? Why bother? *The Language Teacher* 21/5.

Bamford, J. and Day, R. 1998. *Extensive Reading in the Second Language Classroom*. Germany: Cambridge University Press.

Boos, D. and Hughes-Oliver, J. 2000. How Large Does n Have to be for Z and t Intervals? *The American Statistician* 54/2: 121-128.

Bosman, A. and Janssen, M. 2017. Differential relationships between language skills and working memory in Turkish-Dutch and native-Dutch first-graders from low-income families. *Reading and Writing* 30/9: 1945-1964.

Bruton, A. 2009. The Vocabulary Knowledge Scale: A Critical Analysis. *Language Assessment Quarterly* 6/4: 28-297.

Cho, K. and Krashen, S. 1994. Acquisition of Vocabulary from the Sweet Valley Kids Series: Adult ESL Acquisition. *Journal of Reading* 37: 662-667.

Coady, J. and Nation, I. S. P. 1988. Vocabulary and reading. In R. Carter and M. McCarthy (eds.), *Vocabulary and Language Teaching,* London: Longman: 97-110.

Cohen, J. 1992. A power primer. *Psychol Bull* 112/1: 155-9.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. New York: Routledge.

Dabaghi, A. and Rafiee, M. 2012. Incidental Vocabulary Learning and the Development of Receptive and Productive Vocabulary: How Gloss Types Work. *Suvremena Lingvistika* 38/74: 175-187.

Elley, W. and Mangubhai, F. 1981a. *The impact of a book flood in Fiji primary schools*. Wellington: NZCER.

Ellis, N. (ed.) 1994. *Implicit and Explicit Learning of Languages*. San Diego: Academic Press.

Ferreira, M., Almeida, R., and Luiz, R. 2013. A new indicator for the measurement of change with ordinal scores. *Qual Life Res* 22: 1999-2003.

Gaultier, L. 1839. Methode pour entendre grammaticalement la langue latine. In L. G. Kelly, 1969, *25 Centuries of Language Teaching,* Rowley: Newbury House.

Graves, M. 2000. A vocabulary program to complement and bolster a middle-grade comprehension program. In B. M. Taylor, M. F. Graves, and P. Van Den Broek (eds.), *Reading for meaning: Fostering comprehension in the middle grades*, New York: Teachers College Press: 116-135.

Guttman, L. 1944. A basis for scaling qualitative data. *American Sociological Review* 9: 139-150.

Haastrup, K. and B. Henriksen. 1998. Vocabulary acquisition: from partial to precise understanding. In K. Haastrup and A. Viberg (eds.), *Perspectives on Lexical Acquisition in Second Languages*, Lund: University of Lund: 97-126.

Haspelmath, M. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45/1: 31-80.

Hafiz, F. and Tudor, I. 1990. Graded readers as an input medium in L2 learning. *System* 18/1: 31-42.

He, M. 2014. Extensive Reading and Students' Academic Achievement: A Case Study. In T. Muller, J. Adamson, P. S. Brown, and S. Herder (eds.), *Exploring EFL Fluency in Asia*, Palgrave Macmillan UK: 231-243.

Henriksen, B. 1999. Three Dimensions of Vocabulary Development. *Studies in Second Language Acquisition* 21: 303-317.

Hirsh, D. and Nation, I. S. P. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language* 8/2: 689-696.

Horst, M. 2005. Learning L2 Vocabulary Through Extensive Reading: A Measurement Study. *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 61/3: 355-382.

Huckin, T. and Coady, J. 1999. Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition* 21/2: 181-193.

James, M. 1996. Improving Second Language Reading Comprehension: A Computer-Assisted Vocabulary Development Approach. [PhD Thesis]. Honolulu: University of Hawaii.

Jamieson, S. 2004. Likert scales: how to (ab)use them. *Medical Education* 38/12: 1212-1218.

Kelly, L. 1969. *25 Centuries of Language Teaching*. NY: Newbury House Publishers.

Knight, S. 1994. Dictionary Use While Reading: The Effects on Comprehension and Vocabulary Acquisition for Students of Different Verbal Abilities. *The Modern Language Journal* 78/3: 285-299.

Krashen, S. 1989. We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *The Modern Language Journal* 73/4: 440-64.

Krithikadatta J. 2014. Normal distribution. *J Conserv Dent* 17/1: 96-97.

Lai, F. 1993. The effect of a summer reading course on reading and writing skills. *System* 21/1: 87-100.

Laufer, B. 1998. The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics* 19.

Laufer, B. 1989. What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordman (eds.), *Special language: From humans thinking to thinking machines*, Bristol: Multiligual Matters: 316-323.

Law, R. 1991. How many words has the student really learnt? A research into evaluation of long-term vocabulary retention. [MA Thesis]. London: Birkbeck College.

Lix, L., Keselman, J., and Keselman, H. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Rev. Educ. Res.* 66: 579-619.

MacFarland, T. and Yates, J. 2016. Mann–Whitney U Test. In T. W. MacFarland and J. M. Yates (eds.), *Introduction to Nonparametric Statistics for the Biological Sciences Using R*, Springer International Publishing: 103-102.

McLean, S. and Kramer, B. 2015. The Creation of a New Vocabulary Levels Test. *Shiken* 19/2: 1-11.

Meara, P. and Rodriguez Sanchez, I. 1993. Matrix models of vocabulary acquisition: an empirical assessment. In M. Wesche and T. S. Paribakht (eds.), *Symposium on Vocabulary Research*, Ottawa: CREAL.

Meara, P. 1996. The Vocabulary Knowledge Framework. [Unpublished Article]. Accessed at: https://www.lognostics.co.uk/vlibrary/meara1996c.pdf [April 20th, 2021]

Meara, P. and Rodríguez Sánchez, I. 2001. A methodology for evaluating the effectiveness of vocabulary treatments. In M. Bax, and J. W. Zwart (eds.), *Reflections on Language and Language Learning*, Amsterdam: Benjamins: 267-278.

Meganathan, P., Yap, N., Paramasivam, S., and Jalaluddin, I. 2019. Incidental and Intentional Learning of Vocabulary among Young ESL Learners. *3L The Southeast Asian Journal of English Language Studies* 25: 51-67.

Melka, F. 1997. Receptive vs. Productive Aspects of Vocabulary. In N. Schmitt and M. McCarthy (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, Cambridge: Cambridge University Press: 84-102.

Middel, B. and van Sonderen, E. 2002. Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *International journal of integrated care* 2/e15.

Myers, G. 1914. A Comparative Study of Recognition and Recall. *Psychological Review* 21: 442-456.

Nagy, W. 1997. Receptive vs. productive aspects of vocabulary. In N. Schmitt and M. McCarthy (eds.), *Vocabulary Description, Acquisition and Pedagogy*, Cambridge: Cambridge University Press: 64-83.

Nagy, W., Anderson, R., and Herman, P. 1987. Learning Word Meanings From Context During Normal Reading. *American Educational Research Journal* 24/2: 237-70.

Nakanishi, T. 2015. A meta-analysis of extensive reading research. *TESOL Quarterly* 49: 6-37.

Nation, I. S. P. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. 2016. *Making and Using Word Lists for Language Learning and Testing*. Amsterdam: John Benjamins.

Nation, I. S. P. 1990. *Teaching and learning vocabulary*. Boston: Heinle and Heinle.

Nation, I. S. P. 2009. *Teaching ESL/EFL reading and writing*. New York: Routledge.

Nation, I. S. P. 2005. Teaching vocabulary. *Asian EFL Journal* 7/3.

Nation, I. S. P. 2017. The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. Available from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Oxford, R. and Crookall, D. 1990. Vocabulary Learning: A Critical Analysis of Techniques. *TESL Canada Journal* 7.

Palmer, H. 1917. *The scientific study and teaching of languages*. London: Harrap. (Reissued in 1968 by Oxford University Press).

Paribakht, T. and Wesche, M. 1993. Reading Comprehension and Second Language Development in a Comprehension-Based ESL Program. *TESL Canada Journal* 11/1: 9-29.

Pitts, M., White, H., and Krashen, S. 1989. Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language* 5/2: 271-75.

Ponting, A. Pseudo-word List. Accessed [08/25/2021]: https://www.adamponting.com/pseudo-word-list/

Pulido, D. 2004. The effect of cultural familiarity on incidental vocabulary acquisition through reading. *The Reading Matrix* 4/2.

Sarle, W. 1995. Measurement theory: Frequently asked questions. *Disseminations of the International Statistical Applications Institute (Volume 1, 4th Edition)*. Wichita: ACG Press: 61-66.

Schmitt, N. 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D., and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing* 18/1: 55-88.

Schouten-Van Parreren, C. 1996. *Vocabulary Learning and Metacognition*. In K. Sajavaara and C. Fairweather (eds.), *Approaches to Second Language Acquisition,* Jyvaskyla: University of Jyvaskyla: 63-69.

Sinclair, J. 1987. The History of COBUILD. London: Collins COBUILD. Accessed at: https://collins.co.uk/pages/elt-cobuild-reference-the-history-of-cobuild [April 20th, 2021]

Suk, N. 2017. The effects of extensive reading on reading comprehension, reading rate, and vocabulary. *Reading Research Quarterly* 52/1: 73-89.

Suzuki, T. and Sogawa, K. 2010. A Case Study of Motivation Raising Instruction for Successful Learners. *Heles Journal* 10: 3-17.

Tan, D., Pandian, A., and Jaganathan, P. 2016. Lexical Testing and the Reliability of the Modified Vocabulary Knowledge Scale. *Advances in Language and Literary Studies* 7/5: 91-96.

Townsend, J. and Ashby, F. 1984. Measurement scales and statistics. *Psychological Bulletin* 96/2: 394-401.

Waring, R. 1999. *Tasks for assessing second language receptive and productive vocabulary*. [Unpublished PhD Thesis]. Swansea: University of Wales.

Waring, R. and McLean, S. 2015. Exploration of the Core and Variable Dimensions of Extensive Reading Research and Pedagogy. *Reading in a Foreign Language* 27/1: 160-167.

Waring, R. and Nation, I. S. P. 2004. Second Language Reading and Incidental Vocabulary Learning. In D. Albrechtsen, K. Haastrup, and B. Henriksen (eds.), *Angles Volume IV: Writing and Vocabulary in Foreign Language Acquisition,* Copenhagen: University of Copenhagen.

Waring, R. and Takaki, M. 2003. At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language* 15: 130-163.

Webb, S. and Chang, A. 2015a. How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition* 37: 651-675.

Webb, S. and Chang, A. 2015b. Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research* 19/6: 667-686.

Wesche, M. and T. Paribakht. 1996. Assessing vocabulary knowledge: depth vs breadth. *Canadian Modern Language Review* 53/1: 13-40.

Wilkins, D. 1974. *Linguistics in Language Teaching*. Australia: Edward Arnold: 111-112.

Wolfe, H. 1886. Untersuchung Uber Das Tongedachtniss. *Philosophical Studies* 3: 534-571.

Woodrow, L. 2014. *Writing about Quantitative Research in Applied Linguistics*. Palgrave Macmillan: London.

Zahar, R., Cobb, T., and Spada, N. 2001. Acquiring vocabulary through reading: effects of frequency and contextual richness. *Canadian Modern Language Review* 57/3: 541-572.

Zhang, S., Phu, J., Xu, P., Wang, H., Kalloniatis, M., and Zangerl, B. 2021. The performance and confidence of clinicians in training in the analysis of ophthalmic images within a work-integrated teaching model. *Ophthalmic Physiol Opt*. 41/4: 768-781.

Zimmerman, C. 1997. Do reading and interactive vocabulary instruction make a difference? An empirical study. *TESOL Quarterly* 31/1: 121-40.

Xreading URL: https://xreading.com/

Appendix I - The Questionnaire Word List

Core words (n=89):
row
retire
reserve
pity
distance
disappear
blind
blonde
shelf
fail
research
junior
shine
cheer
escape
quality
plane
introduce
pour
awake
official
decision
hire
president
refrigerator
fairy
effort
crowd
clever
curious
example
cloud
alive
exam
success
ghost
deserve
upset
yell
nervous
opportunity
swallow
deliver
attack
receive
law
purpose
guard
sword
weed
article
weak
guest
shock
bark
pleasant
estate
exchange
iron
theater
disappoint
style
teenage
career
traffic
fancy
thin
twin
section
patient
narrow
envelope
biscuit
competition
storm
idiot
tip
breathe
borrow
connect
design
path
improve
sock
immediate
lawyer
chop
roof
rude

Easy word controls (n=5):
play
blue
mother
apple
big

Pseudo-word controls (n=5):
everitual
recide
sweel
gract
reagazine

Appendix II - Single Student SRT Matrix Examples

A. A Single SRT Matrix for Participant 3

	(→)State I	(→)State II	(→)State III	(→)State IV
State I	5	2	1	1
State II	2	4	5	3
State III	2	3	9	7
State IV	0	2	10	38

T1->T2 state changes in receptive vocabulary knowledge for ER- participant 3.

B. A Single SRT Matrix for Participant 24

	(→)State I	(→)State II	(→)State III	(→)State IV
State I	0	1	1	1
State II	0	1	3	2
State III	0	1	9	6
State IV	0	2	23	44

T1->T2 state changes in receptive vocabulary knowledge for ER+ participant 24.

Appendix III - Descriptive Statistics of the SRT Ratings

A. Descriptive Statistics of All Groups and Time Points for Receptive Knowledge

Er- PrQ (n=13)					Er- PoQ (n=13)			
Participant	Mean	Median	Mode	Std. Dev	Mean	Median	Mode	Std. Dev
1	3.28	4	4	1.24	3.89	4	4	0.54
2	3.18	3	4	0.95	2.95	3	4	0.98
3	3.19	4	4	1.02	3.21	4	4	0.99
4	2.96	3	3	0.79	3.68	4	4	0.61
5	3.36	3	4	0.67	3.47	4	4	0.73
6	3.73	4	4	0.82	3.94	4	4	0.35
7	2.80	3	3	0.87	3.09	3	3	0.58
8	2.84	3	3	0.61	2.84	3	3	0.66
9	3.20	4	4	1.10	3.18	4	4	1.10
10	3.77	4	4	0.61	3.86	4	4	0.60
11	3.43	4	4	1.06	3.34	4	4	0.92
12	3.28	4	4	0.94	3.43	4	4	0.71
13	2.50	3	3	0.84	3.30	3	4	0.77
ER+ PrQ (n=12)					**ER+ PoQ (n=12)**			
Participant	Mean	Median	Mode	Std. Dev	Mean	Median	Mode	Std. Dev
14	3.77	4	4	0.78	3.93	4	4	0.39
15	3.65	4	4	0.73	3.78	4	4	0.53
16	3.73	4	4	0.59	3.99	4	4	0.10
17	3.60	4	4	0.85	3.88	4	4	0.38
18	3.40	4	4	0.85	3.95	4	4	0.27
19	2.90	3	3	0.79	3.33	4	4	0.81
20	2.85	3	3	0.69	3.57	4	4	0.63
21	3.84	4	4	0.40	3.57	4	4	0.54
22	3.35	4	4	0.98	3.89	4	4	0.40
23	2.73	3	4	1.15	3.03	4	4	1.20
24	3.61	4	4	0.75	3.51	4	4	0.60
25	3.20	3	3	0.58	3.91	4	4	0.32

The mean, median, mode, and standard deviation of the receptive vocabulary knowledge responses (N=25). Abbreviations: PrQ=Pre-questionaire, PoQ=Post-questionnaire.

B. Descriptive Statistics of All Groups and Time Points for Productive Knowledge

ER- PrQ (n=13)					ER- PoQ (n=13)			
Participant	Average	Median	Mode	Std. Dev	Average	Median	Mode	Std. Dev
1	2.34	2	1	1.23	3.30	4	4	0.99
2	2.96	3	4	1.09	2.91	3	3	0.98
3	2.41	2	2	1.01	2.36	2	2	1.04
4	1.66	2	1	0.76	2.54	2	2	0.70
5	3.07	3	4	0.82	2.80	2	2	0.89
6	3.47	4	4	0.81	2.61	2	2	0.93
7	2.24	2	2	0.80	2.17	2	2	0.46
8	2.59	3	3	0.71	2.78	3	3	0.83
9	2.61	3	4	1.31	2.77	3	4	1.16
10	3.68	4	4	0.72	3.71	4	4	0.78
11	3.21	4	4	1.18	2.87	3	3	1.02
12	2.68	3	3	1.09	2.72	3	3	0.91
13	2.33	2	3	0.87	2.24	2	3	0.85
ER+ PrQ (n=12)					**ER+ PoQ (n=12)**			
Participant	Average	Median	Mode	Std. Dev	Average	Median	Mode	Std. Dev
14	3.52	4	4	0.99	3.86	4	4	0.56
15	3.19	3	4	0.92	3.35	4	4	0.77
16	3.13	3	4	1.00	3.87	4	4	0.45
17	2.80	3	3	0.82	2.83	3	2	0.85
18	2.47	2	3	1.01	3.16	3	4	0.94
19	2.56	3	3	0.81	3.24	3	4	0.85
20	2.65	3	3	0.85	3.29	3	4	0.80
21	3.29	3	3	0.62	3.43	4	4	0.66
22	2.30	2	1	1.08	3.20	3	3	0.71
23	1.83	2	1	0.92	2.54	2	4	1.18
24	2.95	3	4	0.93	2.95	3	3	0.63
25	2.74	3	3	0.97	2.89	3	2	0.87

The mean, median, mode, and standard deviation of the productive vocabulary knowledge responses (N=25). Abbreviations: PrQ=Pre-questionaire, PoQ=Post-questionnaire.

Appendix IV - ANCOVA Test Results

A. Receptive Vocabulary Knowledge ANCOVA on Post-Questionnaire Ratings

Tests of Between-Subjects Effects

Dependent Variable: PoQ

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1.486[a]	2	.743	10.334	.001	.484
Intercept	.888	1	.888	12.347	.002	.359
PrQ	.931	1	.931	12.948	.002	.370
Group	.216	1	.216	3.011	.097	.120
Error	1.581	22	.072			
Total	316.508	25				
Corrected Total	3.067	24				

a. R Squared = .484 (Adjusted R Squared = .438)

The results of a one-way ANCOVA on the ER- and ER+ post-questionnaire mean ratings for receptive vocabulary knowledge with pre-questionnaire data as a covariant.

B. Productive Vocabulary Knowledge ANCOVA on Post-Questionnaire Ratings

Tests of Between-Subjects Effects

Dependent Variable: PoQ

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2.835[a]	2	1.418	13.606	.000	.553
Intercept	2.077	1	2.077	19.937	.000	.475
PreQ	1.485	1	1.485	14.250	.001	.393
Group	1.142	1	1.142	10.959	.003	.332
Error	2.292	22	.104			
Total	226.567	25				
Corrected Total	5.128	24				

a. R Squared = .553 (Adjusted R Squared = .512)

The results of a one-way ANCOVA on the ER- and ER+ post-questionnaire mean ratings for productive vocabulary knowledge with pre-questionnaire data as a covariant.

Appendix V - Mann-Whitney U Tests Comparing I_TC Values Between Groups

A. Receptive Mann-Whitney U Test on I_TC Values of the ER- and ER+ Groups

Independent-Samples Mann-Whitney U Test Summary

Total N	25
Mann-Whitney U	48.000
Wilcoxon W	139.000
Test Statistic	48.000
Standard Error	18.385
Standardized Test Statistic	-1.632
Asymptotic Sig.(2-sided test)	0.103
Exact Sig.(2-sided test)	0.110

The results of a Mann-Whitney U test investigating differences between the ER- and ER+ receptive I_TC data sets.

B. Productive Mann-Whitney U Test on I_TC Values of the ER- and ER+ Groups

Independent-Samples Mann-Whitney U Test Summary

Total N	25
Mann-Whitney U	127.000
Wilcoxon W	205.000
Test Statistic	127.000
Standard Error	18.385
Standardized Test Statistic	2.665
Asymptotic Sig.(2-sided test)	0.008
Exact Sig.(2-sided test)	0.007

The results of a Mann-Whitney U test investigating differences between the ER- and ER+ productive I_TC data sets.