Thresholds, Text Coverage, Vocabulary Size, and Reading Comprehension in Applied Linguistics

by

Myq Larson

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Applied Linguistics.

Victoria University of Wellington
2017
Abstract

The inextricable link between vocabulary knowledge and reading comprehension is incontrovertible. However, questions remain regarding the nature of the interaction. One question which remains unresolved is whether there is an optimum text coverage, or ratio of known to unknown words in a text, such that any deleterious effects of the unknown words on reading comprehension are minimised. A related question is what vocabulary size would a reader need to have in order to achieve the optimum text coverage for a given text or class of texts.

This thesis addresses these questions in three ways. First, a replication and expansion of a key study (Hu & Nation, 2000) was performed. In that study, 98% text coverage was found to be optimal for adequate reading comprehension of short fiction texts when reading for pleasure. To replicate that study, equivalent measures of reading comprehension were collected from a more homogeneous group of participants at a university in northern Thailand (n = 138), under stricter conditions and random assignment to one of three text coverage conditions, to verify the generalisability of the results. The original study was also expanded by measuring reader characteristics thought to contribute to reading comprehension, such as vocabulary size, L1 and L2 literacy, and reading attitudes, in an effort to improve the explainable reading comprehension variance.

In order to more accurately calculate the text coverage a reader experiences for a particular text, both the vocabulary profile of the text and the vocabulary size of the reader must be known as precisely as possible. Therefore, to contribute to the question of vocabulary size, changes such as measuring item completion time and varying the order of item presentation were made to the VST (P. Nation & Beglar, 2007) to improve its sensitivity and accuracy. This may ultimately lead to increased precision when using text coverage to predict reading comprehension.

Finally, L2 English vocabulary size norms were established to supplement the diagnostic usefulness of the VST. Data were collected through an online version of the VST.

---

1Hu and Nation (2000) is occasionally cited as Hsueh-Chao & Nation (e.g. in Brantmeier, 2004) due to a typographical error in the original publication.
created for this thesis from primarily self-selected participants \((n \approx 1.31 \times 10^5)\) located in countries \((n \approx 100)\) around the world representing several L1 and age groups.

Analysis of the data collected for this thesis suggest that text coverage explains much less reading comprehension variance than previously reported while vocabulary size may be a more powerful predictor. An internal replication of Hu and Nation (2000) found errors in the calculation of optimum text coverage and in the reported size of the effect on reading comprehension. A critical review of the theoretical foundations of the text coverage model of reading comprehension found serious flaws in construct operationalisation and research design. Due to these flaws, most research which has purported to measure the effect of text coverage on reading comprehension actually measured the effect of an intervening variable: readers’ vocabulary size.

Vocabulary size norms derived from data collected through an online version of the VST appear to be reliable and representative. Varying item presentation order appears to increase test sensitivity. Despite a moderate effect for L1 English users, item completion time does not seem to account for any variance in vocabulary size scores for L2 English learners.

Based on the finding that vocabulary size may explain both reading comprehension and text coverage, the putative power of text coverage to predict reading comprehension is challenged. However, an alternative measure which may offer greater power to predict reading comprehension, the VST, has been modified and made available online. This version of the VST may provide greater sensitivity and ease of use than the offline, paper-based version.
Dedication

For Professor Thomas Michael Cobb,
for openly sharing with the world a bountiful array of hand-crafted, web-based resources which have long been a source of inspiration. And for taking the time to thoroughly and vigorously critique this thesis.
Acknowledgements

This thesis would not have been possible without various forms of support, advice, and encouragement graciously provided by the following people, organisations, programmes, and things:

General
- Paul Nation
- 邱麗靜 (/tʂʰiɤʊ li.tʂɨŋ/, Li-ching Chiu)
- ธัญญาภรณ์ ผ่องผิว (/tʰan.jaː.pʰɔn pʰɔ̀ːŋ.pʰǐu/, Tanyapon Phongphio)

Resources and finances
- PhD Scholarship awarded by Victoria University of Wellington
- Faculty Research Grants awarded by the Faculty of Humanities and Social Sciences
- Innumerable resources and services donated by Paul Nation

Data
- 138 anonymous participants in Thailand
- ≈ 1.31 × 10^8 anonymous participants around the world

Technical
- Patrick Hindmarsh
- Members of SWEN 302:
  - Cameron Gray
  - Josephine Hall
  - Patrick Hindmarsh
  - Nicholas Lim
  - Daniel Park
  - Felix (Fuci) Shi
  - Philip Walkerdine
  - Henry Williams
- the authors of and contributors to the software listed in Appendix I (page 237)
Morals, ethics, and motivation

- Atheist Community of Austin
- Richard Dawkins
- Daniel Dennett
- Neil deGrasse Tyson
- Sam Harris
- Christopher Hitchens (1949–2011)
- Paul Nation
- James Randi
- Carl Sagan (1934–1996)
- Michael Shermer
- Peter Singer

Miscellaneous

- Every researcher and author cited in this thesis
- Many other individuals and organisations too numerous to itemise
- Coffee (lots)
Preface

Citations

One convention in this thesis may appear, at first glance, to be an error. Careful readers are hereby warned that the citation format in this thesis strictly follows the “Publication Manual of the American Psychological Association” (American Psychological Association, 2010). The consequence of this policy is that the names of some authors appear to be inconsistently cited. Fear not; there is method in the madness. Inconsistencies in self-attribution on the part of some authors in their publications is the cause. The “Publication Manual of the American Psychological Association” provides no latitude to unilaterally normalise citations and doing so would run afoul of the express purpose of a citation: to provide enough accurate and complete information for the reader to retrieve a copy of the cited source (see p. 180). A long history of carelessness in reporting research, detailed in chapter 5 (page 111), gives further justification for presenting authors’ names in the exact format as they appear in publication.

Nevertheless, it useful to protect the sanity of the mindful reader by briefly documenting here the list of authors who, despite variations in their names, are actually the same persons:


Regrettably, there is but one exception to this rule. The first author of Hu and Nation (2000) is listed as Marcella Hu Hsueh-chao in print. Almost every secondary source which cites this study lists the authors as Hu and Nation, Brantmeier (2004) being a notable and admirable exception, so to not follow suit would result in greater confusion.
This minor violation of correcting the typographical error in the original publication still adheres to goal of helping the reader retrieve the original source. This exception is further supported by examining the unpublished Master’s thesis (Hu, 1999) on which the published article is based. The author of that work is listed as Hsueh-chao Marcella Hu.

Non-English text

Another idiosyncrasy which may surprise observant readers is the use of languages other than English when it does not appear to be strictly necessary. This is done for three reasons. The first is practical. Often times there are several methods of transliterating other languages into the Latin-based English orthography. Sometimes, none of the available methods are ideal due to the lack of direct correspondence between the other language and English. Therefore, writing languages in their own orthography is the most accurate representation possible.

On a deeper level, another reason to use native orthographies is to show respect to the study participants and their L1s. Wherever possible, non-English words are written in their respective orthographies followed by both a phonetic transcription for readers who may want to know the pronunciation and an English translation to provide absolute clarity.

Finally, this thesis deals with reading in a L2. Using non-English orthographies explicitly demonstrates to the reader how orthographic distance from a reader’s L1 can present a unique challenge when learning to read in some languages. When encountered, readers are invited to linger on these examples while pondering the challenges of reading in another language.

Research imitates life

This thesis represents a long journey. It is a tale of discovery, disappointment, astonishment, concern, and eventual enlightenment. The journey did not progress according to well-laid plans, necessitating a tortuous and troubling detour through theory, philosophy, and epistemology, finally arriving at a wholly surprising and unexpected destination where all original assumptions and observations are now questioned and viewed with suspicion.

\*\*\*As every reader of English should be well aware, even the current English orthography is not well suited to transparently representing the breadth of English phonology.*
Due to the nature of the journey, the format of this thesis does not follow the traditional structure of an extensive literature review followed by a series of experiments progressively refining a theory or hypothesis. Rather, it presents to the reader the evolution of ideas as they unfolded in a rather chaotic, circular manner. It demonstrates the convoluted path which must occasionally be followed, sometimes returning again and again to first principles, to make sense of the data, disentangle observations from theory, and delineate reality from assumptions.

The information presented in chapter 5 (page 111) examines at great length the historical contexts and technical details of research which had been influential on the formation of currently held beliefs. This is necessary because the conclusions of this critical review are applicable as much to the epistemology of applied linguistics as they are to the specific theory under investigation.

The justification for this unorthodox ordering of ideas is grounded in the idea that researchers should try their utmost to avoid the pitfalls of confirmation bias and approach all experimental results with an open mind. The conclusions drawn are by no means intended to represent the end of the journey nor a pronouncement of absolute truth. Rather it is hoped that the most promising paths for future research will become more apparent and that those which are attractive, but ultimately fruitless, will be precluded from further consideration.
Nomenclature

Terms

1.1  First or native language(s)
1.2  Second or foreign language(s)

running words  The total number of tokens in a text

text coverage  Primarily used after 1970, ratio of word types attested in a particular
word frequency list to running words in a text; cf. vocabulary burden

token  A string of written symbols which represent a single lexical entity similar to the
conventional, non-technical definition of a written word

type/token ratio  Ratio of types to tokens in a text

vocabulary burden  Often used figuratively, or synonymously with readability; primarily
before 1970, ratio of word types unfamiliar or unknown to a reader to running
words in a text; cf. text coverage

(word) type  A unique token within a given text

word  A minimal lexical entity

word family  A grouping of word types which share a common base word type; family
members can be derived through the application of regular morphological pro-
cesses

Initialisms and acronyms

ACTFL  American Council on the Teaching of Foreign Languages
ANOVA  ANalysis Of VAriance
ASRA  Adult Survey of Reading Attitude (M. C. Smith, 1990)
BNC  British National Corpus
DRP  Degrees of Reading Power
EAS  Estes Attitude Scales (Estes, 1971)
ERAS  Elementary Reading Attitudes Survey (McKenna & Kear, 1990)
HTML  Hypertext Markup Language
HTTP  Hypertext Transfer Protocol
IELTS  International English Language Testing System
IP  Internet Protocol
IRT  Item Response Theory
ISO  International Organization for Standardization
MCQ  Multiple Choice Question
NAAL  National Assessment of Adult Literacy
NAEP  National Assessment of Educational Progress
PII  Personally Identifiable Information
PHP  PHP:Hypertext Preprocessor
RSRAA  Rhody Secondary Reading Attitude Assessment (Tullock-Rhody & Alexander, 1980)
SAT  Scholastic Aptitude Test
SD  Standard Deviation (σ)
TF×IDF  Term Frequency × Inverse Document Frequency
TOEFL  Test Of English as a Foreign Language
TOEIC  Test Of English for International Communication
VLT  Vocabulary Levels Test (P. Nation, 1983; N. Schmitt, Schmitt, & Clapham, 2001)
VST  Vocabulary Size Test (P. Nation & Beglar, 2007)
WPM  Words Per Minute
Contents

Abstract i

Dedication v

Acknowledgements vii

Preface ix

Nomenclature xiii

1 Introduction 9
  1.1 Background ........................................... 11
  1.2 Structure ........................................... 12

2 Literature review 15
  2.1 Ranking the usefulness of words ................. 16
  2.2 Optimising the rate of vocabulary acquisition ... 18
  2.3 Ranking the difficulty of texts ................ 25
  2.4 Determining a vocabulary size for reading ... 29
  2.5 Predicting reading comprehension with text coverage ... 31
  2.6 Replication justification ............................... 37
    2.7.1 Participant heterogeneity ......................... 42
      2.7.1.1 Age and L2 acquisition history ............ 42
      2.7.1.2 Gender ........................................ 43
      2.7.1.3 L1 backgrounds ............................... 44
    2.7.2 Instruments ........................................ 45
      2.7.2.1 L1 literacy .................................... 45
      2.7.2.2 Reading comprehension ....................... 46
# CONTENTS

2.7.2.3 Consistency of pseudowords ........................................ 47
2.7.3 Construct validity ......................................................... 48
2.8 Conclusion ................................................................. 49

3 Replication & expansion of Hu and Nation (2000) .................. 59
3.1 Research questions ......................................................... 59
3.2 Participants ................................................................. 62
  3.2.1 Class and age .......................................................... 63
  3.2.2 Gender ................................................................. 63
  3.2.3 Linguistic landscape .................................................. 66
  3.2.4 L2 English acquisition history ....................................... 67
  3.2.5 Literacy ............................................................... 69
3.3 Materials ................................................................. 69
  3.3.1 Experimental text ...................................................... 69
  3.3.2 Pseudowords .......................................................... 72
  3.3.3 Basic measures of reading comprehension ......................... 72
  3.3.4 Expanded measures of reading comprehension .................... 78
  3.3.5 Literacy and reading attitudes ....................................... 79
  3.3.6 Vocabulary size ..................................................... 82
3.4 Procedure ............................................................... 84
3.5 Analysis ............................................................... 85
  3.5.1 Question items from Hu and Nation (2000) ......................... 86
  3.5.2 Expanded study ...................................................... 90
3.6 Results and discussion ................................................... 93
  3.6.1 Limitations .......................................................... 95

4 Internal replication of Hu (1999) ......................................... 97
4.1 Research questions ......................................................... 97
4.2 Original data evaluation .................................................. 99
4.3 Confidence intervals and effect sizes .................................... 101
  4.3.1 Testing assumptions ................................................. 103
4.4 Error checking ............................................................ 108
4.5 Conclusion ............................................................... 110

5 Critical evaluation of text coverage ..................................... 111
5.1 Possible explanations .................................................... 112
5.2 Summaries of previous research ......................................... 115
6 Measuring vocabulary size
6.1 Challenges in measuring vocabulary size 149
6.2 Limitations of the VST 150
6.2.1 Item ordering 150
6.2.2 Completion time 151
6.2.3 Test-taking strategies 152
6.2.4 Over- and under-testing 153
6.3 Research Questions 154
6.4 Refining the VST 156
6.4.1 Item presentation order 156
6.4.2 Decision time 158
6.4.3 Immutable decisions 159
6.5 Participants 162
6.6 Results and analysis 162
6.6.1 Detecting legitimate responses 162
6.6.2 Item ordering effects 165
6.6.3 The time factor 165
6.6.4 Learner norms 167
6.7 Limitations 168
6.8 Conclusion 171

7 Conclusion 173
7.1 Contribution to the field 176
7.2 Limitations 178
7.3 Future research ...................................................... 178
7.4 Implications .......................................................... 180

A Changes to *The Escaped Lunatic* .............................. 181

B VST localised for Thai .............................................. 183

C Literacy Questionnaire ............................................. 189

D Experimental Texts .................................................. 193

E Reading Comprehension Measures ............................... 203
   E.1 MCQ Reading Comprehension Measures .................. 203
   E.2 Constructed Response Reading Comprehension Measures 203

F Human Ethics Committee Approval ............................... 207

G Participant comments on the VST ............................... 209

H Vocabulary Size Norms ............................................. 215

I Reproducible research .............................................. 237

Bibliography .......................................................... 239
# List of Figures

2.1 Hu and Nation (2000) citations ................................................. 38
3.1 Example of VST localised to Standard Thai ............................... 83
3.2 Timeline of data collection .................................................. 84
3.3 Distribution of MCQ scores .................................................. 85
3.4 Model fit evaluation plots for Strict VST Criterion ..................... 87
3.5 Model fit evaluation plots for Loose VST Criterion ................. 89
4.1 MCQ regression plot with added confidence intervals ............ 102
4.2 Cued recall regression plot with added confidence intervals ........ 102
4.3 Revised Hu (1999) model fit evaluation plots ...................... 106
4.4 Revised Hu (1999) MCQ reading comprehension regression .... 107
5.1 Lexical profile comparison between a reader and *Escaped Lunatic* .... 135
5.2 Probable overlap of vocabulary in a reader and *Escaped Lunatic* ........ 135
6.1 VST item presentation order probability density ................... 157
6.2 Median decision time distribution for correct items ............ 164
H.1 L1 English speaker norms .................................................. 216
H.2 L1 Russian speaker norms .................................................. 218
H.3 L1 Chinese (Mandarin) speaker norms .................................. 220
H.4 L1 Japanese speaker norms .................................................. 222
H.5 L1 Arabic speaker norms .................................................. 224
H.6 L1 Vietnamese speaker norms ............................................ 226
H.7 L1 Indonesian speaker norms ............................................ 228
H.8 L1 Spanish speaker norms .................................................. 230
H.9 L1 German speaker norms .................................................. 232
H.10 L1 Korean speaker norms ................................................. 234
LIST OF FIGURES
# List of Tables

2.1 Inconsistent pseudoword replacement in Hu & P. Nation (2000) ........................................ 47
2.2 Carver's (1994) five reading processes ........................................................................... 49
2.3 Summary of indicative text coverage research ................................................................. 51
2.3 Summary of indicative text coverage research (continued) ............................................. 52
2.3 Summary of indicative text coverage research (continued) ............................................. 53
2.3 Summary of indicative text coverage research (continued) ............................................. 54
2.3 Summary of indicative text coverage research (continued) ............................................. 55
2.3 Summary of indicative text coverage research (continued) ............................................. 56
2.3 Summary of indicative text coverage research (continued) ............................................. 57
2.3 Summary of indicative text coverage research (continued) ............................................. 58
3.1 Vocabulary size and comprehension correlations in Hu (1999) ........................................ 62
3.2 Distribution of treatment and class by gender and age. ..................................................... 64
3.3 Participant L2 English learning history ............................................................................ 68
3.4 Hours reading per week in L1 and L2 ............................................................................. 70
3.5 Number of pseudowords per text coverage level ............................................................ 71
3.6 Reading comprehension variance, Strict VST Criterion .................................................. 87
3.7 Reading comprehension variance, Loose VST Criterion .................................................. 88
3.8 Linear regression of reading comprehension on text coverage ........................................ 89
3.9 Descriptive statistics for expanded predictor variables. ................................................... 91
3.10 Variances of expanded predictor variables ..................................................................... 92
3.11 Fitted model .................................................................................................................... 93
4.1 Comparison of original and replication summary statistics ............................................. 100
4.2 Reading comprehension regression with confidence intervals ....................................... 101
4.3 Variances of Hu (1999) independent variables ............................................................... 104
4.4 Comparison of original and revised MCQ regressions ..................................................... 106
4.5 Calculation differences for real and discrete values ......................................................... 109
5.1 Reading comprehension operationalisation comparison .......................... 128
5.2 Reading comprehension thresholds comparison ................................. 131
5.3 Tokens and coverage for *Lady Chatterley’s Lover* .......................... 134

6.1 Original item proportions in online VST groupings ............................ 157
6.2 Participant L1S ............................................................................. 160
6.3 Location of participants .................................................................. 161
6.4 Median correct response decision time comparison ............................ 164
6.5 Vocabulary size and decision time comparisons ................................. 166

A.1 Canges to *The Escaped Lunatic* ................................................... 182

H.1 Vocabulary Size Test norms for L1 English users ............................. 217
H.2 Vocabulary Size Test norms for L1 Russian users ............................ 219
H.3 Vocabulary Size Test norms for L1 Chinese (Mandarin) users .......... 221
H.4 Vocabulary Size Test norms for L1 Japanese users .......................... 223
H.5 Vocabulary Size Test norms for L1 Arabic users .............................. 225
H.6 Vocabulary Size Test norms for L1 Vietnamese users ..................... 227
H.7 Vocabulary Size Test norms for L1 Indonesian users ...................... 229
H.8 Vocabulary Size Test norms for L1 Spanish users ............................ 231
H.9 Vocabulary Size Test norms for L1 German users ........................... 233
H.10 Vocabulary Size Test norms for L1 Korean users ............................ 235
Chapter 1

Introduction

Dr. West has prepared an edition of Stevenson’s *Treasure Island* that requires a reading vocabulary of only 1779 words. Dr. West mentioned this edition in an address at Columbia University a year or two ago and someone [sic] in the audience arose to criticize him for so “mutilating” an English classic.


When learning to read in a l2, a conflict arises for all but the youngest of learners: a reader’s l2 reading ability is almost always lower than that of someone who happens to share the same birth year but is an l1 user of that language. The shared ages of this hypothetical learner and the l1 user pair make it highly probably that they will share nearly identical levels of cognitive development. This odd juxtaposition of linguistic immaturity and mental sophistication on the part of learner leads to difficulty in sourcing l2 material which is appropriate to the reader.

Two options are available to the learner: read texts which are appropriate for the readers’ cognitive age while accepting that the text is likely to be linguistically challenging, perhaps prohibitively so if the difference in age and ability is great, or read texts which are linguistically appropriate for the readers’ l2 ability while enduring content that is likely more appropriate for less cognitively mature readers. Neither option is ideal and the severity of the situation is determined by the degree of tolerance on the part of the reader to the mismatch. Some readers might overlook insipid content to im-
prove L2 reading ability while others may be willing to struggle through a dense network of unknown words and unfamiliar syntax in search of meaningful ideas.

This is an oversimplification of the problem; content and linguistic complexity are not actually as strongly correlated as that but the problem of matching readers with L2 texts is a real challenge. Any method which purports to address it will require accurate measurements of reader characteristics which are indicative of, at minimum, generalised reading ability and L2 proficiency. Also required are accurate measurements of text characteristics which correlate with, or causally determine, the levels of reader characteristics required to comprehend the text.

Skilled authors have an intuitive sense of which text characteristics are important and can tailor their texts to match the average reader characteristics of their intended audience. When that audience consists primarily of L1 users of the language in which the text is written, authors can use a combination of age and education level to estimate those reader characteristics with reasonable precision.

Texts which are written for an audience of L2 readers present unique challenges because the correlation between linguistic ability and cognitive development is much weaker. Although some positive correlation probably does exist, neither age nor education level can be sufficiently precise predictors of L2 ability to be useful to authors because L2 acquisition can commence at any stage in a language learner’s life and the trajectory of L2 development varies much more widely than L1 development. Though daunting, this is a challenge which cannot be overlooked.

Reading is an important activity in L2 acquisition (Grabe, 2004; Krashen, 1989). This should be uncontroversial. The general agreement regarding the efficacy of reading in supporting L2 development has led to the establishment of extensive reading programmes in many L2 teaching institutions throughout the world. One of the biggest challenges to L2 reading programmes is not in sourcing material. Rather, the challenge lies in accurately recommending material which is at an appropriate level for the reader.

These programmes are predicated on the belief that their greatest benefit can be best realised by maximising L2 input through reading. Furthermore, unless a reader is appropriately matched to an L2 reading text, several desirable outcomes may not be achieved. Reading speed may be reduced, for example, which precludes the opportunity to maximise input. The reader’s ability to comprehend basic factual propositions encoded in the text may be impaired if the linguistic complexity is too challenging. Or opportunities to acquire new vocabulary, both intentionally and incidentally, may be sacrificed if the text is not sufficiently rich in lexis. Ignoring very important reader characteristics such as background knowledge and motivation and text characteristics such as genre, this
thesis presents research investigating the lexical aspects of matching readers to text. The research and analysis contained within this thesis should help make the matching process more precise and robust by testing one common vocabulary-focused matching method and by establishing vocabulary size norms which can be used in that and more diverse methods.

1.1 Background

Researchers have been quantifying vocabulary (e.g. Thorndike, 1921; Kaeding, 1898) and texts (e.g. C. L. Johnson, 1927; Lively & Pressey, 1923) for more than a century. A wide range of research themes have arisen from this basic activity, each differing slightly in the problems addressed and interests of the researchers. Often times the overlap between them is greater than their differences. Making frequency lists has perhaps the longest tradition, with a history measured in centuries (e.g. Bright, 1588/1888) and continues to this day (e.g. Michel et al., 2011). This type of research is foundational to many other research themes. Studies which search for the optimal rate of acquiring the words on those lists (e.g. M. West, 1926/2003) constitutes another research theme. Another uses the words in those lists to rank texts in terms of difficulty (e.g. C. L. Johnson, 1927) to create a sequence of progressively more lexically challenging texts. Still another research theme attempts to determine the number of words on the lists which need to be learnt in order to accomplish certain reading tasks (e.g. I. S. P. Nation, 2006; Hirsh & Nation, 1992).

Matching readers to texts is a task which benefits greatly from these research themes. One method in particular can be thought of as a close sibling: the use of text coverage to predict reading comprehension (Laufer, 2013). This method involves estimating the overlap between the reader’s mental lexicon and the lexical profile of a text. This can not be done directly, so a reader’s vocabulary size is estimated using a vocabulary size test and the contents are modelled with a word frequency list. The same word frequency list is used to create a lexical profile of a text. Once the overlap reaches a specific threshold, usually around 95% (Laufer, 1989) or 98% (Hu & Nation, 2000), the reader is predicted to have a high likelihood of reading the text with an adequate level of reading comprehension.

For this method to be useful, a predictable relationship must exist between text coverage and reading comprehension. Yet the effectiveness of this method is questionable.

---

1Hu and Nation (2000) is occasionally cited as Hsueh-Chao & Nation (e.g. in Brantmeier, 2004) due to a typographical error in the original publication.
Studies suggest that one or more text coverage levels represent one or more thresholds, above which readers are more likely than not to achieve a specific degree of reading comprehension (e.g. Laufer, 2013; N. Schmitt, Jiang, & Grabe, 2011). But these studies differ in the constructs measured and conflate text coverage with vocabulary size. The causal mechanisms of text coverage are also unclear. Some researchers theorise that text coverage is a threshold which prevents a reader from drawing on L1 reading ability when engaging in L2 reading (e.g. Laufer & Sim, 1985). Another suggestion is that the effect might be located directly in the text itself, blocking direct access to the text’s meaning (see R. Anderson & Freebody, 1981). These uncertainties persist largely because there have been few studies which experimentally isolate text coverage when measuring reading comprehension.

1.2 Structure

This thesis addressed this problem directly. The first study (chapter 3, page 59) investigated the relationship between text coverage and reading comprehension through an experimental research design modelled on the only prior study to employ a similar design (Hu & Nation, 2000). As all other studies have been observational, the question of causality remained an open question. Culturally and linguistically homogeneous participants were asked to read and answer questions about manipulated texts in which specific proportions of the running words were replaced with pseudowords to ensure specific text coverage levels. The participants’ L1 and L2 literacy, gender, and L2 vocabulary sizes were also measured as possible additional explanatory variables.

After that experiment produced only very weak confirmatory results, the second study (chapter 4, page 97) was conducted in the form of an internal replication of Hu and Nation (2000). Summary data was obtained and reanalysed in an attempt to explain the differences with the first study. Uncovering significant errors which cast doubt on subsequent research, the theoretical framework of the text coverage model of reading comprehension was then subjected to a meticulous critical review (chapter 5, page 111).

As a result of the increased suspicion of the text coverage model, the final study (chapter 6, page 147) departed sharply from the trajectory implied by the first study. It sought to refine methods of measuring vocabulary size and establish age- and L1-based norms, both of which are useful to not only the questionable text coverage model of reading comprehension, but also other methods of matching readers to texts. The data for the final study were collected from thousands of participants around the world via an online version of the test.
Experimentally investigating and subsequently critically challenging the validity of the text coverage model is an important step in the development of more reliable methods of matching readers to texts. Progress is made in this regard through continually testing and comparing the skillfullness of the predictive models and retaining only the most promising. Likewise, establishing the normed vocabulary size ranges for various readers is also an important step in improving the accuracy of the methods of matching readers to texts as well as a wide variety of other L2 acquisition research tasks where quantifying participant L2 proficiency is important.
Chapter 2

Literature review

In an letter to the editor of Science, Kirkpatrick (1891) detailed a novel methodology for estimating the vocabulary size of “ordinary” (p. 107) people. It involves first selecting a text which could be easily understood by all readers. In this case, the text chosen was De Foe’s (1850) Robinson Crusoe. Then, by tallying the unique word types that occurred in the novel, a lower bound estimate of the reader’s vocabulary size could be established by making the assumption that most of the word types in the novel were known to the reader. By estimating the lower bound, i.e. a minimum vocabulary size, which could reasonably be assumed of a person performing the reading task, Kirkpatrick avoided the problem of overestimation to which competing vocabulary size methodologies were prone.

Although Kirkpatrick’s primary goal was to advance vocabulary size methodology, the unique approach provides an insight into the prevalent view at that time regarding the relationship between vocabulary and reading. In discussing the validity of the new methodology, Kirkpatrick conceded that young readers wouldn’t know all of the word types, but that they would “probably” (p. 108) recognise 90% when reading Robinson Crusoe for pleasure. That caveat encapsulates the crux of Kirkpatrick’s methodology: no

1Although most modern orthographic conventions have settled on Defoe as the canonical spelling for the author of Robinson Crusoe, the author used both De Foe and Defoe as an adopted surname. Kirkpatrick (1891) likely used the former spelling because contemporaneous editions of Robinson Crusoe (e.g. De Foe, 1850) listed the author’s name as Daniel De Foe on the title page. The biographical memoir about the author printed in some of editions of that era noted that De was a prefix which the author adopted after being born “a plain Foe” (De Foe, 1862, p. xv).

2The tally was an estimate arrived at by extrapolating from pages systematically sampled from the text. This was necessary due to the labour-intensive process the technique demanded. However, Kirkpatrick’s figure of “not less than five to six thousand words” (p. 108) is surprisingly accurate given the methodology. A modern computer analysis indicates the text comprises approximately 6100 word types.
one really knew what percentage of the word types in a text were actually known when reading for pleasure. To compensate for this unknown, Kirkpatrick simply fudged the final estimate, in sharp contrast to the precision that the methodology promised to bring to the question. Without an accurate figure though, Kirkpatrick’s manual adjustment to the word type count would result in an overestimation of the reader’s vocabulary size which, ironically, the methodology sought to avoid.

Thus is the morass one falls into when researching the relationship between vocabulary and reading. There is a long history of research on this topic replete with myriad unknowns, implicit assumptions, and avenues of research which simultaneously bring clarity to particular questions yet oftentimes cloud the larger picture as a whole. This chapter will attempt to delineate some of the prominent research themes from the beginning of the 20th century to the present time while acknowledging that such divisions are inherently artificial and shortsighted. The significant overlap in the motivations of previous researchers intricately bind their studies together thematically, but it is imperative to impose boundaries in order to properly situate the experiment detailed in chapter 3 (page 59).

2.1 Ranking the usefulness of words

Although Kirkpatrick’s (1891) methodology was unique, it is representative of a body of research at that time which focused on counting, ranking, and categorising words. Despite lacking the computing power that modern day corpus linguists take for granted, the close of the 19th century through the first few decades of the 20th century was a period of intense effort to statistically describe human language.

Characterising words in terms of frequency was justified, in part, by one of four broad curriculum reform movements in the United States of America which attempted to increase educational efficiency and minimise waste in much the same fashion that factories had done during the previous decades of industrialisation (Kliebard, 2004). The movement promoted the notion that knowledge could be transmitted more efficiently than traditional education methodologies, which focused primarily on the humanities, when it was compartmentalised, ordered, structured appropriately. A related goal of the efficiency movement was exclude from the curriculum skills and topics which were unlikely to be useful to a learner’s future. According to Kliebard, three concrete outcomes of the movement were:
1. the establishment of junior high schools which gave educators time to establish what future potential each learner had and, based on that determination, which educational pathway would be most appropriate in high school;

2. a shift from teaching the subject of history as a collection of facts about the past to a focus on more directly practical subjects such as citizenship and social studies;

3. the teaching of reading based on science and word frequency lists.

As one of the most basic building blocks of reading is the humble word, the influence of the efficiency reform movement on the field of reading, at least in the United States of America, is completely understandable. Given the ideals of the movement, it makes little sense to spend limited classroom time teaching the meaning of rare words which are unlikely to be encountered in daily life while sacrificing the opportunity to teach high frequency words which, by definition, are regularly encountered.

The zeal with which this goal was pursued is evident in that period’s research environment, one “increasingly dominated by a torrent of scientific studies of word frequency.” (Kliebard, 2004, p. 128) Within a relatively short time span, myriad word frequency lists were published such as Thorndike’s (1921) *The Teacher’s Word Book*, initially published with a listing of 10,000 English words, expanded to 20,000 words a decade later (1931), and expanded yet again to 30,000 words the next decade (1944). Word frequency lists for other languages were also widely published during that era, such as for Latin (Browne, 1907), French (Henmon, 1924), Spanish (Buchanan, 1927), and German (Morgan, 1928; Kaeding, 1898).3 The construction of these lists continue to the present day with modern analogues such as Zeno, Ivens, Millard, and Duvvuri’s (1995) *The educator’s word*

3According to Buchanan (1927), initial interest in word frequency lists was initiated by the field of stenography in the 1800s when impressively large corpora were created to characterise German (e.g. Kaeding, 1898). But the idea of identifying the most common words in order to optimise some language-related task can be traced back at least as far as Bright (1588/1888) who created one of the first recorded shorthand systems for English. Bright’s system replaced words with symbols, superficially similar to an idiographic or symbolic writing system such as Chinese. However, Bright noted that Chinese

[...] chara\textcyr{t}ers are very long, and harde to make, that a dou\textcyr{t}en of mine, may be written as foone as one of theirs; Befides, they wanting an alphabet, fal into an infinite number, which is a thing that greatlie chargeth memory, and may discourge the learner. (p. A4)

Clearly Bright intended to keep the repertoire of symbols small and consequentially limit the number of possible words they could represent. This implies that word frequency was a criterion which informed his selection of meaning-bearing words. Phonological neighbourhood density and synonymy were also likely criteria as Bright’s system used the same symbol to represent near- and exact-homophones as well as their synonyms.
frequency guide as well the ever increasing long list of frequency lists produced from large corpora such as Leech, Rayson, and Wilson’s (2001) list, Word Frequencies in Written and Spoken English, tabulated from the venerable BNC. The Google corpus of millions of books and billions of tokens (Michel et al., 2011) takes the field to a level that would have been inconceivable to the pioneers of this line of research.

Frequency lists were and are created for other purposes as well, but their existence and the social forces which helped bring them into existence in the first instance brilliantly illustrate this first research theme, one focused on reducing waste and improving efficiency in education. Thorndike (1921) noted that the lists represented a “measure of the importance of words” (p. iii), although imperfect, and recommended several ways that the measure could be useful to teachers in improving the efficiency of their lessons. Of course, teaching learners to read was the ultimate goal, but the link between the vocabulary research and reading in this theme is the weakest when compared with the other themes described in this chapter. Some of the basic assumptions supporting this research theme, on which the proceeding themes will build, include the following:

- frequency in a corpus is a reasonable proxy of a word’s usefulness to a learner;
- more frequent words need to be taught to learners, i.e. they are unlikely to be learnt incidentally or already known;
- the return on investment of teaching and learning vocabulary is proportional to frequency.

As with each thematic research summary in this chapter, the list of assumptions is neither an indictment nor commendation of the theme and the research which comprise it, but rather an attempt to bring conscious awareness and clarity to each theme’s most salient features.

2.2 Optimising the rate of vocabulary acquisition

If the assumptions supporting the ordering of vocabulary by frequency are valid, the next research theme which brings vocabulary and reading even closer together is one which attempts to determine the most efficient rate of acquisition. After all, it would be antithetical to the movement to determine the rank order of word usefulness but then neglect teaching that list at the most optimal rate.

Perhaps best known in the L2 vocabulary research community for A General Service List of English Words, a reference book which lists not only the 2000 most frequent words in a five million word corpus of English, but also the frequency of the meanings of each of the words, M. West’s (1953) research best embodies this research theme. While the
list itself has left an indelible mark on the field of applied linguistics, is considered by some (e.g. Gilner, 2011) to be one of the quintessential symbols of the vocabulary control movement, and also helped propel M. West’s status to that of “one of the world leaders of English language teaching” (Howatt, 1984, p. 249), it also clearly has its roots in the frequency lists which had already been published. M. West’s intended use of the list was to assist in the optimisation of learning by providing a principled foundation for implementing an ideal rate of vocabulary acquisition through reading.

M. West’s body of research in L2 studies is more than just a single, well-known publication, though; it is voluminous and spans more than forty years (M. West, 1968). The ideas and opinions on vocabulary expressed in these works, although dated, are still afforded a higher platform of credibility, whether justified or not, than many of M. West’s contemporaries. Given the lingering influence and continuing high citation rates, it is useful to consider M. West’s philosophy and influences broadly to place these influential works in the correct context.4

As an educator in the “difficult circumstances” (M. P. West, 1960) of the Bengal region located in present day India and Bangladesh, the prevalence of inefficiencies in the education system coloured M. West’s view on language learning, resulting in a strongly pragmatic hue. These experiences and observations are summarised in Learning to read a foreign language (1926) which is primarily a summary of a larger, more extensive monograph published in the same year: Bilingualism (1926/2003). In the latter, M. West described a detailed and extensive needs assessment which revealed, among other things, that many pupils attended school irregularly and for only a few years. This made it practically impossible to plan a curriculum which continuously built upon content studied over many years. The efficacy of teaching English was abysmally low and consumed an indefensible amount of time and resources which could more profitably be spent on other topics more likely to be useful to a typical Bengali boy.5 In general, time invested in teaching English, both as a foreign language and as a medium of instruction, was squandered with disappointing results (M. West, 1953, chapter 5).

Much of M. West’s work can be viewed as a reaction to the inefficiency of teaching and the lacklustre results it achieved, but it also comports with the zeitgeist of the education efficiency reforms at the beginning of the 20th century in the United States of America (Kliebard, 2004). As a practical solution to this situation, M. West sought to

---

4 Howatt (1984) is an excellent and informative account of the history of L2 English teaching which includes an entire chapter devoted to M. West’s ideas and influence.

5 Girls and schools for girls are only rarely mentioned in M. West’s (1926/2003) Bilingualism but there is no reason to believe that the results would not be equally applicable to Bengali girls.
empower students, as quickly and efficiently as possible, with enough reading ability in
English that they could potentially use that knowledge as a bootstrap to continue both
educating themselves after their time in the formal education sector came to a close as
well as have the opportunity to make up for any deficiencies in that education system
(Howatt, 1984).

Based on the experience gained from preparing controlled vocabulary reading texts
for use in the experiments detailed in Bilingualism, M. West went on to create The New
Method Readers series along with a short pamphlet (M. West, 1927) detailing the process
by which further texts could be created in line with the criteria M. West proposed. Al-
though consideration was given to a variety of design goals for the readers, the most
germane to this research theme is the balance M. West sought to achieve between creat-
ing context-rich texts which were enjoyable to read and texts which were conducive to
vocabulary acquisition. It is a difficult balance to achieve. For M. West, a reader should
not be interrupted by encountering unknown vocabulary too often lest reading becomes
a chore, but no interruptions would result in no improvement in vocabulary knowledge.
In finding the ideal compromise, M. West decided that

[s]ix new words per page would be the highest limit at which the child could
preserve the continuity of the story and enjoy his task. We may therefore
decide that New words should not be introduced more frequently than one new
in every fifty running words of the text. (M. West, 1926, p. 27, emphasis in
original)

In other words, 98% of the running words should be known to the reader while just 2%, or
one word in fifty, should be introduced as new word types. This particular vocabulary
burden, or the ratio of unfamiliar words to running words in a text (Carr, 1934), is a
figure which many researchers have adopted (e.g. Carr, 1934; P. Nation & Coady, 1988).
Carr (1934), for example, called for the creation of 12 Latin readers to assist learners
in acquiring vocabulary after noting that children learning to read in l1 English were
reading texts with the same vocabulary burden of 1 : 50 and that French readers had
also been created with a vocabulary burned of 1 : 56.

Despite existing evidence demonstrating the superior efficacy of context-free paired
associate learning (e.g. Thorndike, 1908; Ebbinghaus, 1913), a widely held belief con-
tinued to justify vocabulary acquisition through reading: new words are best learnt in
context.6 This belief can be clearly seen in Weigel (1919) who noted that, in the processes
of vocabulary acquisition,

6Firth (1968) is famous for noting that context is an essential aspect of comprehending the meaning
of the word, as well as defining the concept of collocation, by writing “[y]ou shall know a word by the
[a] word only then gains its true significance when it stands in relation to other words. In fact, it has now this, now that meaning, depending upon its relations. Hence to recall a word, the easiest way is to lay up a store of sentences in which it is used. (p. 351)

The belief was so pervasive at the time that M. West (1926) enshrined it as one of the axioms of teaching L2 reading: “[n]ew words should be learnt by practice in actual reading situations, not memorised as vocabularies” (M. West, 1926, p. 27; M. West, 1926/2003, p. 269). In the early decades of the 20th century, learning words in context, primarily through reading with occasional reference to a dictionary (Lodge, 1907), seems to have been the more widely followed belief, at least amongst L2 researchers and practitioners, than paired associate learning.

It is interesting to note that this ideal vocabulary burden for optimal vocabulary acquisition is probably higher than what Kirkpatrick assumed L1 English readers experienced when reading Robinson Crusoe for pleasure. Kirkpatrick’s (1891) vocabulary size calculations were based on a reader recognising only 90% of the word types in the text. That figure can be used to estimate what the vocabulary burden might have been. The electronic version of the edition likely used by Kirkpatrick7 contains ≈121000 running words, or tokens, and ≈6100 word types. As with almost all texts, the nature of the Zipfian distribution means that ≈100 most frequent word types will account for ≈60% of the running words. Yet 60% of the word types will only occur between one and three times each in the text, accounting for relatively few running words.

Since Kirkpatrick was considering L1 English readers reading an English text, it is probably safe to assume in this estimate that the unknown word types were primarily, if not exclusively, located in the group of word types with a frequency ≤ 3. Just as the frequency of the second most frequent word type in a text is about half of that of the most frequent word type and about double that of the third most frequent word type, the same pattern mostly holds for the count of word types per frequency. For example, the number of word types appearing twice in a text is about half as many as those appearing only once in a text and about double those appearing thrice in a text. Therefore, Zipf’s law can be used to estimate the distribution of the unknown word type frequencies which can then be used to calculate the probable vocabulary burden of Robinson Crusoe.

company it keeps!” (p. 179), but it is clear that Weigel and his contemporaries were cognisant of this notion as well.

7De Foe (1850) can be downloaded from https://books.google.com/books?id=ED_QAAAAMAAJ.
Zipf’s law can be described formally as

\[ f(k; s, N) = \frac{1}{k^s \sum_{n=1}^{N} \left( \frac{1}{n^s} \right)} \]

where \( s \) is an exponent describing the distribution, \( N \) is the number of word type frequencies, and \( f \) is the frequency of a word type frequency of rank \( k \). For human languages, \( s \) is very close to 1 and the number of word types to estimate would be \((1 - .9) \times 6100 = 610\). The number of unknown word types with a frequency of 1 can be estimated as:

\[
610 \times \frac{1}{\sum_{n=1}^{3} \left( \frac{1}{n^s} \right)} \approx 333,
\]

166 with a frequency of 2, and 111 with a frequency of 3.

These estimates can be used to calculate an approximate vocabulary burden as follows:

\[
\text{burden vocabulary} = \frac{t_{\text{unknown}}}{t_{\text{running}}} = \frac{333 \times 1 + 166 \times 2 + 111 \times 3}{121000},
\]

which evaluates to 0.82% or, inverted to form a ratio, the vocabulary burden can be written as 1 : 121, which is more than twice as high as M. West’s (1927) ideal of 1 : 50. Kirkpatrick’s conjecture was intended to describe L1 English readers while M. West was referring to L2 English readers from Bengal. But the difference in relatively small numbers, i.e. unknown words comprising 2% compared with 0.82% of a text, is of immense importance. If the density of new words is too low, then vocabulary acquisition will become slow and inefficient. Too high, however, and the learning burden will become too great, again resulting in an inefficiency. Both of these outcomes would have been the anathema of the scientific-based efficiency reform movement.

There is one glaring difference between graded readers, à la M. West (1932), and \textit{Robinson Crusoe} which must be noted. M. West’s (1932) \textit{The new method readers for teaching English reading to foreign children} are inauthentic, intentionally created texts. Authentic, unmodified texts, and perhaps graded readers created without the frequency lists and principles of construction used by M. West, had a much longer history of use in L2 education. They also continued to be used in educational settings long after M. West (1927) was published, so researchers had little reason to not also evaluate them in terms of their potential to support vocabulary acquisition as well. Perhaps the difference between authentic and contrived texts would explain the differences in the vocabulary burden between the two types?
2.2. OPTIMISING THE RATE OF VOCABULARY ACQUISITION

Some attempts were made to experimentally identify the ideal vocabulary acquisition rate using authentic texts, although with limited success. Holley (1973)\(^8\) noted that there was little empirical support for an ideal vocabulary burden, labelled by Holley as “new-word density” (p. 339), which would facilitate vocabulary acquisition. Hypothesising that there exists a threshold for new-word density that, when passed, will lead to inefficient vocabulary acquisition, Holley used low frequency synonyms which were assumed to be unknown to the participants (\(n = 246\)) to systematically replace between five and 50 high frequency words in a 750 word fable. The procedure yielded seven versions of the fable with new-word densities of 1 : 15, 1 : 18.75, 1 : 25, 1 : 30, 1 : 37.5, 1 : 75, and 1 : 150. The study’s participants were adult L2 German learners with similar L2 German learning histories. One key departure from M. West’s (1926/2003) experiment is that Holley did not rewrite the text following the principles in M. West (1927). Instead, a glossary was provided to facilitate acquisition of the new words in a primarily authentic text, a method described by Carr (1934) as the “‘dictionary-thumbing’ or ‘finger-exercise’ method” (p. 324). No evidence was found to support the hypothesis of a new-word density threshold for vocabulary acquisition. Nor did there appear to be any effect on reading comprehension. These findings led Holley to conclude that any threshold that might exist may well be beyond the range tested in the study and that the field should be wary of “assumptions about foreign language learning without substantive research findings to back them.” (p. 346)

There are serious limitations to Holley’s (1973) study, the most notable being the research design and statistical analysis. The time available to learn the new words was not constant per word; the group tasked with learning just five new words had the same amount of time as the group which tried to learn an order of magnitude more. In that sense, the design measured the rate of acquisition in terms of time, but not so much in terms of new-word density. Additionally, the performance of the groups on the vocabulary test which followed was not normalised despite the large difference in the number of items in each group’s test. Finally, the results were analysed using ANOVA. As the groups did not differ in treatment, but only in degree, a regression analysis would have been more appropriate. The only time ANOVA will result in a significant difference

---

\(^8\)A summary of Holley (1973) alongside other important studies discussed in this chapter can be found in Table 2.3 (page 52).
in this sort of situation is when the researcher spaces the degrees of treatment far enough apart with respect to the effect coefficient.⁹

Despite the lack of strong empirical support for the supremacy of the method, almost two decades later, Krashen (1989) continued to carry the flag for the vocabulary-acquisition-through-reading camp by boldly asserting the validity of the theory in an unambiguously titled study: “We Acquire Vocabulary and Spelling by Reading”. Carver (1994)¹⁰ countered these claims with a large study \((n = 568)\) of young and adult L1 English readers who self-reported unknown vocabulary in a series of short texts which were relatively difficult or easy. The ease of the texts relative to the participants’ ability was determined by the difference between a measure of participant reading ability known as the Accuracy Levels Test and a measure of text difficulty using either the Degrees of Reading Power (DRP) method or the Rauding Scale of Prose Difficulty (Carver, 1990). Carver reported that unknown words only represent about 1% of the running words in a text when its difficulty is matched to the reader’s ability, almost no words are unknown when the text’s difficulty rating is less than the reader’s ability, and about 2% are unknown when the text’s difficulty is greater than the reader’s ability level. These findings suggest that texts would need to be relatively difficult to provide enough new words to explain vocabulary acquisition through reading. The problem for Krashen’s (1989) theory is that readers do not usually read difficult texts (Carver, 1994) and thus never reach the optimal acquisition rate promoted by M. West. This discrepancy between ideals, theory, and research serves as one of the primary motivations for the research described in chapter 3 (page 59). More recent studies dealing with this theme will be described in proceeding sections.

The influence of M. West’s recommended vocabulary burden as an optimal rate for vocabulary acquisition through reading held sway for decades, cemented by M. West’s status, extensive list of publications, and participation in meetings which ultimately brought together the world leaders of the vocabulary control movement and culminated in the publication of *A General Service List of English Words* (Howatt, 1984). The influence of the optimal vocabulary burden figure and, in particular, the theory behind it are still manifest to this day in the wide range of graded readers available to L2 English learners.

Naturally, the overview of M. West’s ideas and influence on others in this research theme is myopic and overly simplified. The debate between the efficacy of intentional

---

³Increasing the number of participants \((n)\) will also usually increase the odds of obtaining a significant result from most statistical tests, but this critique is based on analysing the data that were collected, not what hypothetically could be collected.

¹⁰A summary of Carver (1994) alongside other important studies discussed in this chapter can be found in Table 2.3 (page 54).
paired associate learning compared with reading graded readers such as those advocated by M. West has also continued for decades as the theoretical foundations of educational psychology transitioned from behaviourism to more cognitive models. Roughly a decade after *A General Service List of English Words* was published, for example, Carroll (1964/2000) asserted that paired associate learning was “a better strategy than making the quasi-mystical assumption that vocabulary can ‘automatically’ be acquired by any student who does wide reading.” (p. 16)

Nevertheless, this oversimplification allows for some of the implicit assumptions to come to the surface here. In addition to the assumptions required by the previous research theme, some of the important assumptions to explicitly note in this research theme include:

- an optimal rate of vocabulary acquisition exists;
- the rate is invariant with respect to the learner;
- the rate is invariant with respect to the word type;
- learning new vocabulary items in context is an efficient method of acquiring vocabulary;
- materials can be created which incorporate the optimal rate;
- authentic texts can be accurately characterised in terms of the ratio of high to low frequency words or some similar metric;
- the rate at which the lower frequency words are encountered in authentic texts is similar to the optimal vocabulary acquisition rate.

### 2.3 Ranking the difficulty of texts

This and the following research themes can be viewed as branches that grew from the trunk of frequency lists and optimal learning rates. The tree analogy is apt here because the previous themes comprised research which dealt with the more basic building blocks of language and vocabulary acquisition. They are straightforward tasks which are, *prima facie*, useful. The research in the previous themes probably would have been described at the time of their publication as “serious scientific studies [which] have begun to supplant guesswork in the preparation of reading material for pupils learning to read English and other modern languages at various levels of difficulty.” (Carr, 1934, p. 325)

All of the following research themes rest upon them. Yet the studies which comprise this and the following research themes begin to diverge in focus and application, much like branches from the trunk of a tree. Like branches growing further from the trunk, the
likelihood of overlap and intermingling among them increases despite diverging from the same trunk some time ago. It is for this reason that the artificial distinctions imposed to create these research themes become less tenable, but this endeavour will, hopefully, continue to justify and situate the experiment described in chapter 3 (page 59).

This research theme describes the effort to use the vocabulary in a text to determine the difficulty of a comprehending it. The goal here is to ensure that a text is in the reader’s Goldilocks Zone, i.e. neither too easy or too difficult, in the much the same manner that M. West strove to find a balance between efficient encounters with new vocabulary while maintaining reading pleasure. But this goal is distinct to the previous research theme because the primary emphasis here is to ensure a given text will be understood by the reader. While this was, of course, also one of the criteria considered by M. West (1927), the research here pays little to no interest in the potential vocabulary acquisition benefits of reading.

The origin of this research theme can largely be traced to C. L. Johnson (1927). Rather than searching for the ideal ratio of unknown to known words to optimise learning, C. L. Johnson sought to estimate text difficulty from that ratio instead. This idea is a subtle reversal in the direction of the effect used by Kirkpatrick (1891) to estimate vocabulary size. Whereas Kirkpatrick conjectured that easily understood texts were written with predominantly well-known words, Johnson surmised that texts written with predominantly well-known words would be easily understood. In C. L. Johnson’s own words, “[i]f the difficulty of reading is directly proportional to the difficulty of vocabulary, then we could grade reading texts by determining the relative degree of difficulty of their vocabularies.” (pp. 290–291)

To confirm this correlation, C. L. Johnson\footnote{A summary of C. L. Johnson (1927) alongside other important studies discussed in this chapter can be found in Table 2.3 (page 51).} rated 37 texts based on the prevalence of difficult vocabulary within the texts. Eighteen L2 French learners were then asked to read two of the texts and were subsequently asked to evaluate which of the two was “harder” (p. 294). The percentage of uncommon words in each text ranged from 5% to 16%. In general, the learners’ subjective impressions of relative difficulty were in alignment with C. L. Johnson’s vocabulary-based predictions, thus providing empirical support that this approach to predicting text difficulty had merit.

In that study, uncommon words were not necessarily unknown; they would be called infrequent words today. Texts with more infrequent words would tend to be relatively more difficult than those with fewer. C. L. Johnson’s tentative conclusions and simplicity of calculation seem to had become elevated to the status of absolute certainty in the
minds of some when Carr (1934) confidently proclaimed that “the reading difficulty of a given passage of Latin or of any other language can be quite accurately measured by its vocabulary burden; that is, by the proportion of unfamiliar words to the running words in the passage.” (p. 323, original emphasis)

It is important to take note of the subtle shift in direction from Kirkpatrick (1891) to C. L. Johnson (1927). It is logical fallacy, but not because it is always false; it is just not always true. Recall that Kirkpatrick used simple propositional logic to arrive at a conclusion based on direct reasoning, formally known as modus ponens and written symbolically as \( P \rightarrow Q; P \). The conditional statement, \( P \rightarrow Q \), can be extracted from the following statements:

In order to determine the size of an ordinary vocabulary I could think of no better means than to find out the number of words used in some standard work that is easily read and understood by everybody. […] Children of ten or twelve years read the book with pleasure, and probably have a pretty clear idea of the meaning of nine out of ten of the words they find in it. (Kirkpatrick, 1891, p. 108)

In other words, if a standard work (Robinson Crusoe) is easily read and understood by everybody (\( P \)), then children will know the meaning of 90% of the words contained within (\( Q \)). The entire logical argument, with a generalised conditional statement, can be written as:

1. If a book is easily read by all, then almost all of its words are understood by the reader.
2. Robinson Crusoe is easily read by all.
3. Therefore, almost all of its words must be understood by the reader.

Provided that the conditional statement is true, and intuitively it seems that it should be, then the conclusion must be true if the antecedent is true.

The only valid transformation of this argument is the contrapositive, formally known as modus tollens. To construct the contrapositive of the argument, the last two statements must be swapped and negated as can be seen in the symbolic form: \( \frac{P \rightarrow Q, \neg P}{\neg Q} \). So the valid contrapositive of Kirkpatrick’s argument would be

1. If a book is easily read by all, then almost all of its words are understood by the reader.
2. Almost all of the words in Robinson Crusoe are not easily understood by the reader.
3. Therefore, Robinson Crusoe is not easily read by all.
CHAPTER 2. LITERATURE REVIEW

C. L. Johnson’s argument is an example of affirming the consequent, also known as the fallacy of the converse. It can be written symbolically as $P \rightarrow Q ; P \rightarrow \frac{Q}{P}$ or as an argument:

1. If a book is easily read by all, then almost all of its words are understood by the reader.
2. Almost all of the words in Robinson Crusoe are easily understood by the reader.
3. Therefore, Robinson Crusoe is easily read by all.

Although this argument is only consistently valid if $P$ is the only sufficient condition for $Q$, it can still be true under some conditions. It may seem intuitively uncontroversial that increasing the proportion of difficulty of vocabulary comprising a text will result in a concomitant increase the perceived difficulty of the text, but vocabulary might not be the only factor.

It is the search for such conditions which brought about the first readability formulae. C. L. Johnson was not the first to broach the idea of predicting text difficulty from an algorithm which takes as parameters quantifiable characteristics of the text itself; Lively and Pressey (1923) are credited as the originators (Rader, 1922; DuBay, 2004). But where Lively and Pressey investigated three methods of quantifying the vocabulary load of texts through corpus methods which would today seem primitive, C. L. Johnson was one of the first to confirm the results with actual readers. The focus on vocabulary was perhaps inevitable given the intense interest in quantifying vocabulary at that time, but only a few years passed before readability formulae evolved into more complex and powerful tools.

Gray and Leary (1935) conducted one of the earliest and impressively extensive studies of readability. In an effort to determine the true meaning of the term readable, over 200 factors identified as possibly affecting reading comprehension were considered and measured in about 350 books. Five of the most promising factors were found that could be combined to yield a correlation of $R = .6435 \pm .085$ (p. 138) with reading comprehension. Gray and Leary’s (1935) research spawned dozens of other readability studies which refined the measurements of texts and reading comprehension as well as simplified the formulae. Today, most readability formulae consist of two categories of factor: syntactic complexity and vocabulary difficulty (DuBay, 2004).

These formulae are criticised for being atheoretical (Bernhardt, 2011), which is true because they make no claims of causality (Bormuth, 1966) and are thus only predictive models of reading comprehension. Because they make no comment on causation, their simplicity and high correlations should not cause any “intellectual indigestion” (Bormuth, 1966, p. 129). Nevertheless, there are some additional assumptions in this research theme which should be made explicit here. Unlike the previous research themes,
however, not all of the proceeding themes will incorporate these assumptions, but it is important to clarify them before moving on. Using vocabulary to rank the difficulty of texts assumes that

- texts differ in the level skill required to comprehend them
- some characteristics of the word types in a text affect the ability of a reader to comprehend that text;
- word type frequencies in a corpus is one such characteristic;
- the likelihood that a word type in a text will be known to a reader is proportional to its frequency in a corpus.

Since few of today’s readability formulae rely solely on vocabulary measurements,\(^{12}\) this thesis will not dwell on the prediction of text difficulty. But this research theme has been necessary to review because the research in chapter 3 (page 59) will focus on a special case of this idea.

2.4 Determining a vocabulary size for reading

Before arriving at the research theme which most precisely locates the research in this thesis, one more very closely related research theme must be explored. After word frequency has been characterised, optimal learning rates established, and texts have been ranked for acquisition optimisation and/or difficulty, a question then arises as to the terminus of the vocabulary acquisition journey which the title of a relatively recent study, I. S. P. Nation (2006), puts succinctly: “How large a vocabulary is needed for reading and listening?”

Compared with the preceding research themes, the question here is slightly more distant from the vocabulary acquisition focus of Carr (1934) and M. West (1927). Learning vocabulary through reading is not the focus, although clearly vocabulary must be acquired by some method in order to read. The method of answering the question in this research theme hearkens back, as so many of the research themes do, to Kirkpatrick (1891) and the assumption that readers will know most of the word types when reading *Robinson Crusoe*. Conventional wisdom of that time, according to an unspecified source mentioned in Kirkpatrick (1891), was that general reading and conversation required a knowledge of 1000 to 3000 words but Kirkpatrick’s calculations suggested readers of *Robinson Crusoe* would need to know roughly double that figure. What could explain

\(^{12}\)McLaughlin’s (1969) SMOG Grading is probably the most notable exception. The only parameter this readability formula takes is the number of $\geq 3$ syllabic words per 30 representative sentences in a text.
the discrepancy? Which figure would be a better goal for a learners? What vocabulary size would one need to have in order to read all other novels similar to Robinson Crusoe?

The research in this theme approaches the question of vocabulary size in much the same manner as C. L. Johnson (1927) who theorised that texts comprised of mostly common words would be easily understood: “the higher the coverage [of known words], the more likely the text will be understood.” (Webb & Macalister, 2013, p. 303) If that is true, one simply needs to calculate the number of word types in a corpus of representative texts, not unlike Thorndike and Lorge’s (1944) word frequency lists, and adjust that figure by the highest acceptable vocabulary burden when reading. This is easier said than done.

It is not clear what the highest acceptable vocabulary burden would be. Perhaps it is around 1 : 50, 98% known running words, similar to the ideal vocabulary burden for promoting vocabulary acquisition through reading. Maybe it is lower, around 1 : 121, or 0.82% known running words, as Kirkpatrick assumed. Is it higher? D. B. Johnson (1972) noted that a vocabulary burden as high as 1 : 11, or 91% known running words, might be acceptable after finding that a number of foreign language word frequently lists all shared similar lengths of between 4000 and 6000 word types, represented the number of word types needed for “liberated” (p. 125) reading according to the experience of language teachers, and provided ≈ 91% coverage of representative texts.

Despite the dreams and ideals of the educational efficiency movement, and the Herculean efforts required to manually tabulate the frequency lists used to evaluate texts for their vocabulary learning potential, empirical support for the maximum vocabulary burden tolerated by readers was lacking. Whatever the correct figure is, which will be discussed in the next research theme, there is no shortage of studies which attempt to calculate recommended vocabulary sizes based on it. In I. S. P. Nation (2006), for example, a variety of vocabulary size recommendations are determined by performing a vocabulary profile analysis on several small corpora comprising texts representative of various genres. The recommended sizes range from 8000 to 9000 word families for written texts and 6000 to 7000 word families for spoken texts. Around 10,000 word families were found by Webb and Macalister (2013) to be required for books written for younger L1 users as well as ones written for older L1 users. However, Webb and Rodgers (2009) make recommendations ranging from 3000 to 10,000 for comprehending movies.

13 This idea is a fallacy of the converse and thus not guaranteed to always be true, but is not always false, either.
14 This is an estimate. See section 2.2 (page 22) for details.
15 D. B. Johnson’s (1972) use of the term “liberated” reading is actually taken from Scherer (1965).
2.5. PREDICTING READING COMPREHENSION WITH TEXT COVERAGE

The extremely large range is due to uncertainty over the most acceptable vocabulary burden. It is important to determine the figure with as much accuracy as possible because the vocabulary size recommendations emanating from the studies in this research theme are based on it. As noted by N. Schmitt et al. (2011), “[t]he key factor in these widely varying estimates is the percentage of vocabulary in a text that one needs to comprehend it.” (p. 26)

Each of the preceding research themes has progressively incorporated a greater bond with successful or adequate reading comprehension and this theme is no different. It should be clear that this research theme is even more intimately related to reading comprehension. The vocabulary burden figure used to derive the vocabulary size estimates is set to the highest point possible before reading comprehension is impaired. The assumptions needed to support this theme incorporate the assumptions of the preceding research themes and incorporates the following:

- measurement and characterisation of the mental lexicon is possible;
- the skills needed to read unmodified texts are relatively invariant with respect to text type or genre;
- one of the primary skills needed to read unmodified texts is
  - a vocabulary size which is at least a minimum size,
  - knowledge of a specific list of “useful” words,
  - or knowledge of a specific percentage of running words in a text.

2.5 Predicting reading comprehension with text coverage

This final research theme is the one to which this thesis belongs. It deals precisely with the question of the proportion of running words which need to be known in order to read a text for pleasure with reasonable comprehension. These studies focus on the effect of unknown vocabulary on reading comprehension more directly than in any of the preceding research themes. Determining a vocabulary size for reading, described in the preceding research theme, crucially relies on the accuracy of this figure, so no matter what the current estimate of the figure is, there is always scope for improvement. Reducing confidence intervals of almost every important estimate should be viewed as a worthy endeavour, especially when the figure is used as an input in solving real world problems. These incremental increases in the precision of measurement are one of the hallmarks of scientific advancement.
Studies which attempt to predict reading comprehension from the proportion of unknown vocabulary in a text can not be fully appreciated without reference to three key hypotheses, all of which are derived from the concept known as cross-linguistic transfer. In general, transfer refers to influences, both positive and negative, between languages during the acquisition process and forms one of the primary theoretical assumptions of bilingual education (Garcia, 2000). When applied to the case of L2 reading, transfer can refer to both metalinguistic transfer, such as the use of cognates to work out the meaning of unknown words in a text, or cognitive and metacognitive strategies, such as the use of background knowledge and comprehension monitoring in the process of reading comprehension.

Cummins’s Interdependence Hypothesis (1979) and Threshold Hypothesis (1976) are perhaps the most influential transfer-based hypotheses. The Interdependence Hypothesis asserts that “[t]o the extent that instruction in Lx is effective in promoting proficiency in Lx, transfer of this proficiency to Ly will occur provided there is adequate exposure to Ly (either in school or environment) and adequate motivation to learn Ly.” (Cummins, 1981, p. 29) In other words, there are generalised, language-independent cognitive skills whose development in any language can be beneficial when a learner is using other languages. The Threshold Hypothesis posits that there exists some linguistic thresholds which must be surpassed before generalised, language-independent cognitive skills developed in one language become available in the other. A lower threshold must be met to avoid potential negative academic consequences of being bilingual, and a higher threshold marks the point where proficiency in both languages is high enough for the generalised skills to be equally available for use (Cummins, 1994). Clarke’s (1980) Short Circuit Hypothesis can be viewed as a special case of the Threshold Hypothesis; it states that strong readers will revert to poor reading strategies when the reader’s L2 proficiency is below a threshold.

In both the Short Circuit Hypothesis and the Threshold Hypothesis when applied to reading, the ability that is inhibited by the threshold is the use of higher-order reading abilities and metacognitive strategies. It was this threshold which Laufer and Sim (1985) attempted to describe and quantify by interviewing small groups of L2 English learners until one group was found whose members’ L2 abilities did not inhibit the reading strategies they needed to deal with three academic texts. For all of the groups, none of the participants (n > 12, total n unspecified) were reported to have problems performing the same tasks when reading academic texts in their L1S, so they all presumably had the skills which could be made available in L2 reading once the threshold was crossed. After two interviewers jointly assessed several groups of learners of in-
2.5. PREDICTING READING COMPREHENSION WITH TEXT COVERAGE

creasing L2 English ability, one group \( (n = 7) \), whose proficiency level was described by Laufer and Sim as likely Advanced Plus according to ACTFL Proficiency Guidelines, was found whose members seemed to be able to reasonably “cope with” (p. 408) the academic texts. This was thought to be evidence that the participants’ L2 English ability had exceeded Cummins’ (1976) and Clarke’s (1980) thresholds. The group members could then be characterised by their performance on a recognised standardised test to quantify that threshold. The interviewers observed that all of the participants, including those at the threshold level, struggled with vocabulary and their performance notably improved when the meaning of troublesome vocabulary was provided to them. This factor was felt to have a much larger impact than deficits in subject matter and grammatical knowledge which lead Laufer and Sim to conclude that vocabulary knowledge most strongly determined whether reading skills could transfer for use in L2 English reading.

According to Laufer and Ravenhorst-Kalovski (2010), the first study to propose a link between a specific lexical coverage threshold and reading comprehension was Laufer (1989). Although it should be clear from the preceding research theses that this claim is not entirely accurate, there is some merit in that the concept of an absolute lexical threshold to reading comprehension was absent, or at best only implied, in previous research. It is this specific link which distinguishes the studies in this research theme from the others. It marks a shift in emphasis away from text difficulty to a prediction of reading comprehension performance. This is a critical distinction to make because this idea subtly changes the focus from simply ranking texts in terms of some characteristic to establishing a critical value for reading comprehension analogous to the way the term is used in statistics. This critical value was determined in Laufer (1989) to be 95% of the running words in a text through the following methodology.

First year L1 Hebrew and L1 Arabic university students \( (n = 100) \) read one of two academic texts while underlining unknown words and answering either open-ended or MCQ reading comprehension questions. Participants were then given a translation task to verify the accuracy of the reported unknown words. After the proportion of known to unknown running words, adjusted by the results of the translation task, were determined for each participant, Laufer partitioned the participants into two groups and found that when 95% known running words was used as the critical value for determining group membership, the number of participants whose reading comprehension test performance was greater than 55% became significantly larger the number in the other group. Laufer therefore concluded that knowing 95% of the running words in a text represented a threshold below which “reasonable” (p. 317) reading comprehension would be unlikely.
It should be noted at this point that the term for the proportion of known running words in a text used in Laufer (1989) was not vocabulary burden, but rather text coverage. Around the early 1970s, as computers made it possible to calculate the lexical profiles of entire texts rather than just samples of texts calculated by hand, the concept of text coverage emerged. Similar to vocabulary burden, text coverage refers to the ratio of known to unknown running words in a text. But text coverage also often refers to proportion of running words to those not appearing in a given word list rather than, in the case of vocabulary burden, those assumed to be unknown to the reader. The need for a new term appears to be due to the use of computers to calculate the “coverage”\(^{16}\) (D. B. Johnson, 1972, p. 125) of a text provided by a word frequency list of a particular size. Similar to C. L. Johnson’s (1927) shift in directionality, the shift in definitions here is subtle, but important.

When vocabulary burden was the prevailing paradigm, lists such as Thorndike and Lorge’s (1944) The Teacher’s Word Book of 30,000 Words were consulted to estimate the proportion of running words which were not on that list. Texts with greater proportions of tokens were considered to be more difficult to read, thus requiring a higher level of language proficiency. But the list itself was viewed as more of a syllabus or learning goal rather than a proxy for the reader’s mental lexicon. Word types in a text which were unknown to a reader, but on the list, should be taught as they were viewed as useful while those not on the list could be ignored. In Laufer (1989), the operationalisation of text coverage was essentially the same as vocabulary burden because the readers themselves indicated which words were unknown in the text. However, many studies in this research theme make no distinction between vocabulary burden, as originally conceived, and the overlap between a word list and a lexical profile as mentioned in D. B. Johnson (1972).

In order to determine a causal relationship, though, it is important to maintain this distinction. There have been a number of studies which build upon Laufer (1989) but use text coverage, in the D. B. Johnsonian sense, to infer vocabulary burden indirectly. Laufer and Ravenhorst-Kalovski (2010)\(^{17}\) is probably the most extreme example. In that study, a word frequency list derived from the BNC was used to determine the text coverage of older versions of a reading comprehension test which participants \((n = 745)\) had recently taken. The participants’ vocabulary sizes, estimated using the VLT, were

---

\(^{16}\)D. B. Johnson (1972) appears to be one of the first researchers to use the concept of coverage in this way, apparently borrowing from Steinfeldt (1965), and described it as “a sine qua non in any definition of reading fluency.” (p. 125, emphasis in original)

\(^{17}\)A summary of Laufer and Ravenhorst-Kalovski (2010) alongside other important studies discussed in this chapter can be found in Table 2.3 (page 56).
then substituted for the same number of ranked entries in the word frequency list. As
text coverage was not actually measured directly, and despite the authors’ claim that
“data confirmed some earlier results regarding the percentage of text lexis [...] required
for reading comprehension” (p. 26), little can actually be concluded regarding the vari-
ance on the reading comprehension test explained by text coverage. Another example
of confounding the two is Hirsh and Nation (1992) where frequency lists and lexical
profiles alone were used to determine an ideal text vocabulary burden for L2 readers.
There were no actual participants in that study, yet the ideal vocabulary burden was
determined “[...] by relating vocabulary coverage to ease of reading.” (p. 692)

N. Schmitt et al. (2011) appeared to be cognisant of this problem. To estimate the
vocabulary burden of two texts, participants (n = 661) were asked to indicate know-
ledge of more than 50% of the word types in those texts on a Yes/No test. The participants
later read the texts and answered reading comprehension questions about them. A linear
relationship (r = .407) was found between performance on the reading comprehension
measure and the calculated vocabulary burden. This approach appears promising, and
it is an improvement over the aforementioned studies because fewer assumptions need
to be made about the vocabulary burden experienced by the participants. But some
assumptions must still be made.

Knowledge of the word types not appearing on the Yes/No test were still estimated
by using text coverage. But even if 100% of the word types were tested in advance,
there are still few conclusions that can be drawn from the study. As with Laufer and
Ravenhorst-Kalovski (2010), the study was observational, not experimental. This is why
a regression analysis can not be performed on the data and no predictions can be made.
But for some reason, N. Schmitt et al. do make predictions, asserting that “an increase
in comprehension growth of 2.3% for each 1% growth in vocabulary knowledge” (p. 39).
The basis for making this claim is not specified and no regression coefficients are re-
ported. The other problem with this observational design is that the vocabulary burden
was determined solely by the vocabulary size of each individual participant. Thus, the
vocabulary burden variable was a combined characteristic of the participant and the text.
Without controlling for vocabulary size, it is impossible to know whether any reading
comprehension variance can be attributed to the unknown vocabulary in the texts or
simply the unique vocabulary size of each participant.

A summary of N. Schmitt et al. (2011) alongside other important studies discussed in this chapter can
be found in Table 2.3 (page 58).
In the same year, Prichard and Matsumoto (2011)\(^9\) took a similar approach. Participants \((n = 103)\) were asked to translate into 11 Japanese 71 word types from a 650 word text. It is not clear what proportion of total word types this was, but for a text of that length, it is probably reasonable to assume that knowledge of approximately 25% of the word types was tested; all others were apparently assumed to be known. Unique to this study, only word types which were not translated or incorrectly translated were used to calculate the vocabulary burden experienced by the readers. A modest correlation of \(r = .29\) was reported but, as with other studies, vocabulary burden was determined by the participants’ own vocabulary size, making interpretation of the statistic problematic. Despite that challenge, Prichard and Matsumoto (2011) speculated that knowing 97% of the running words in a text might be enough to achieve adequate reading comprehension.

To address the problem of causality directly, Hu and Nation’s (2000)\(^20\) participants \((n = 66)\) were invited to read one of four versions of the same text followed by a short survey and two types of reading comprehension questions. Low frequency tokens in three of the texts had been replaced with pseudowords, strings of letters which conform to standard orthographic and phonological constraints of a language and are interpreted as plausible but unknown words. By definition, they were unknown to the participants so specific vocabulary burden figures, independent of the participants’ vocabulary size, could be assured. The remaining tokens were within the 2000 most common English word families which Hu and Nation assumed the participants knew well based on their performance on the VLT. This design ensured vocabulary burden alone was the unambiguous explanatory variable. The two reading comprehension measures served as semi-redundant measures of the dependent variable, reading comprehension, while the survey simply measured the participants’ predicted confidence in their performance on said measures.

Based on a linear regression analysis of measured reading comprehension, Hu and Nation concluded that most learners of English could achieve “adequate unassisted comprehension” (p. 422) of fiction texts at 98% vocabulary burden. This conclusion has been met with wide acceptance within the applied linguistics field, boasting citations by a wide range of highly respected academics (e.g. Koda, 2005; N. Schmitt, 2008; Read, 2004; Milton, 2009). Rather than using Laufer’s (1989) recommended vocabulary bur-

\(^{9}\)A summary of Prichard and Matsumoto (2011) alongside other important studies discussed in this chapter can be found in Table 2.3 (page 57).

\(^{20}\)A summary of Hu and Nation (2000) alongside other important studies discussed in this chapter can be found in Table 2.3 (page 55).
den of 95%, for example, I. S. P. Nation uses 98%, citing more recent studies by Hu and Nation (2000) and Carver (1994). Why was the latter figure preferred?

The 98% text coverage figure does have historical precedence (M. West, 1926) but the 98% figure was likely preferred due to the increase in methodological rigour apparent in Hu and Nation (2000). It is one of the few studies which have investigated the relationship between vocabulary burden and reading comprehension using an experimental, rather than an observational, design. This fact might partly explain why Hu and Nation (2000) quickly began to stand out over other text coverage studies and it is the reason why this thesis will not only replicate the core elements of Hu and Nation (2000), but will also expand on it by adding additional measures in an attempt to improve the predictive power of the model.

Before moving on to a more detailed justification for replicating Hu and Nation (2000) and some of its limitations, the assumptions of this final research theme, the last in a series of related research themes which progressively brought closer together the topics of vocabulary knowledge and reading comprehension, can be added to the list:

- the text coverage of a text can be determined with reasonable accuracy;
- the vocabulary size of a reader can be determined with reasonable accuracy;
- text coverage is equivalent with vocabulary burden, i.e. for a given text, the proportion of words in that text not on a frequency list of a specific size is similar to the proportion of words in that text not in the mental lexicon of a reader with the same vocabulary size;
- frequency lists derived from large corpora are reasonably representative of the course of vocabulary acquisition in a typical reader.

### 2.6 Replication justification

Of all the studies investigating the relationship between text coverage and reading comprehension, Hu and Nation (2000) is both highly cited and the only experimental study dealing with L2 English readers. Citations of Hu and Nation (2000) obtained from

---

21A fuller discussion of M. West (1926) can be found in section 2.2 (page 20).
22Although a careful distinction between text coverage and vocabulary burden has been made up to this point, the term text coverage will be used henceforth for two reasons. First, the methodology of Hu and Nation (2000) makes it possible to assume with some confidence that vocabulary burden as experienced by the participants was almost the same as the text coverage calculated by comparing the text’s word types with a frequency list. Second, the term vocabulary burden has fallen out of use even though many of modern text coverage studies are implicitly referring to it. The use of text coverage for the remainder of this thesis implies a truly synonymous relationship with vocabulary burden.
Figure 2.1: Hu and Nation (2000) citations. The line is an ordinary least-squares regression in the form \( a \times e^{b \cdot x} \) where \( a = 3809.13 \) and \( b = -34.01 \).

Google Scholar (https://scholar.google.com) and plotted in Figure 2.1 (page 38) can be viewed as a proxy of the interest in and the popularity of the study.\(^{23}\) It is impossible to measure the value and faith that each citation places on the research without reading each individual paper, but the raw citation data and the sustained rate of citations, occurring at the rate of one citation every 6 days over the past five years, give an impression of the intensity and breadth of discussion of the study within the academic community. The volume of that discussion is on par with research which is well-established within the field and rarely questioned due to its methodological rigour and robust results.

As a basic tenet of epistemology, facts should be thoroughly supported by some sort of evidence. They are labelled as facts because of their enduring nature as evidenced by repeated experience or interaction yielding the same results. The best predictive models in any discipline similarly are only accepted as such after they have been shown to successfully, and repeatedly, predict future events. For Hu and Nation (2000) to be

\(^{23}\)As with any algorithmically-generated data, citation counts reported by Google Scholar are vulnerable to manipulation or error. However, a cursory check of some of the citations indicate that the counts provided by Google Scholar are sound. For example, Figure 2.1 (page 38) shows that Hu and Nation (2000) was cited once, before it was published, in 1999. At first blush, this may appear to be an error, but the article in question (P. Nation & Ming-Tzu, 1999) shares a coauthor and the citation notes that Hu and Nation (2000) is "in progress" (P. Nation & Ming-Tzu, 1999, p. 379).
accepted as an established model of reading comprehension, as the citation rate suggests it is becoming, it must be replicated.

Curiously, most research results, according to Ioannidis (2005b), particularly those produced by smaller studies with few participants, have small effect sizes and follow a non-standardised design, are actually false. Furthermore, even results which are not subsequently found to be outright false are still subject to refinement, especially if the data represent the first measurement of a given phenomenon. For example, out of 34 of the most highly cited studies from top medical journals which found positive effects for interventions and were subsequently replicated, Ioannidis (2005a) found that seven (≈ 21%) were contradicted when replicated and a further seven (≈ 21%) were found to report effects which were stronger than those found in replications. Only 20 (≈ 59%) were replicated. In other words, approximately 41% of the most highly-regarded and widely-cited medical studies were reporting effects which were either exaggerated or flat out wrong. This phenomena is found throughout the social and natural sciences.

There are many reasons why this is the case, but one of the more obvious reasons is appropriate to detail here. Simply put, novel research is somewhat exploratory in nature and therefore does not justify investing the extensive resources necessary to conduct a large-scale study. Therefore, datasets generated by this type of research tend to be relatively small and consequently more prone to yielding spurious results. If a phenomenon exhibits a rather small effect, then small studies of that phenomenon are unlikely to detect any effect at all, and consequently risk never being submitted for publication. But some studies will also detect the rare extreme examples of the effect which would certainly warrant publication. This is known as the File Drawer Problem (Rosenthal, 1979).

The fact that this phenomenon exists is not a bug in the scientific process, but rather, a feature. The only way to ensure that the parameter estimates revealed by research are correct is to verify the results through replication. Replication is the essential, and slightly unglamorous, backbone of the scientific method. It is an inescapable mandate

---

24Goodman and Greenland (2007) assert that the Ioannidis (2005b) study itself is false in a technical sense, and that the proportion of false studies is actually around 14%, but agree with the spirit of Ioannidis (2005b) that many findings are spurious and steps need to be taken to avoid the phenomenon in medical research.

25For accessible overviews of Ioannidis’s research, see Lehrer (2010) and Freedman (2010).
that is born from the understanding that error is present in every measurement, especially when the number of participants\textsuperscript{26} \((n)\) and/or effect size\textsuperscript{27} is small.

Replication studies, even in fields which are stereotypically considered rigorous, are always in short supply. Polio and Gass (1997) note that the field of linguistics, in general, values replication less than in related fields such as psychology which consequentially transfers to some researchers in applied linguistics. This is unfortunate due to the tendency within the field of applied linguistics to “base generalizations on relatively few subjects.” (p. 500) Due to the difficulties in replicating research on human subjects, as opposed to replication studies in the physical sciences which tend to lack ambiguity, they note that virtual replications, where all of the conditions in the original study are identical, are impossible.

According to Mackey (2012), research in applied linguistics should be replicated when a study is “generally accepted in the field, but might have been insufficiently investigated” (p. 28). Based on those criteria, it appears that Hu and Nation (2000) is an ideal candidate for replication. The popularity and appeal of predicting reading comprehension from text coverage, such as in studies which make vocabulary size recommendations based on the model,\textsuperscript{28} rests largely on the claims set out in Hu and Nation (2000) because it is essentially the only experimental study to date.\textsuperscript{29} Outside of this study, though, there is little empirical justification to support the text coverage model of reading comprehension. The vast majority of research corroborating the text coverage model is based on intuition and conjecture (e.g. Hirsh & Nation, 1992) or correlation (e.g. Laufer & Ravenhorst-Kalovski, 2010; N. Schmitt et al., 2011).

\textsuperscript{26}There were 67 participants in Hu and Nation (2000), but pair-wise deletion resulted in only 66 participants being used in the statistical analyses. One participant did not complete the MCQ reading comprehension measure and another did not complete the cued recall reading comprehension measure. List-wise deletion would have resulted in 65 participants.

\textsuperscript{27}For details on the effect size and confidence intervals of Hu and Nation (2000), see section 4.3 (page 101).

\textsuperscript{28}For further details, see section 2.4 (page 29).

\textsuperscript{29}Yoon (2011) is an approximate replication of Hu and Nation (2000). In it, Thai learners of Korean read four short texts at five vocabulary burden levels. The participants completed two reading comprehension questions for each text. The statistical analysis (ANOVA) reported that 24% of the reading comprehension variance could be explained by vocabulary burden. It is unclear, however, why ANOVA was used to compare means rather than characterise the relationship between the two variables. Despite some commendable improvements in methodology, such as the repeated measures, it is difficult to directly compare the results to Hu and Nation (2000). A summary of Yoon (2011) alongside other important studies discussed in this chapter can be found in Table 2.3 (page 57).

Given the need to replicate Hu and Nation (2000), the decision of how to replicate the study must be made for replications are not as simple in applied linguistics as they are in the physical sciences. Polio and Gass (1997), noting this problem, make the distinction between virtual replications, where everything is the same as in the original study, and conceptual replications, where the most salient claims and features are replicated. In practice, it is difficult to conduct a virtual replication in applied linguistics because the participants cannot be described with enough precision to locate equivalent participants for the replication. But conceptual replications may be more appropriate to L2 research anyway since they involve testing the generalisability of claims. The specific details, such as the environment or subjects, may differ to the original, but if the claims are robust, the same effects should be observed. The problem, Polio and Gass (1997) point out, is that should a replication fail to yield the same results, uncovering the source of the discrepancy can be challenging.

Porte’s (2012) taxonomy is perhaps more useful starting with the distinction between internal replications, where data are reanalysed by the original researchers, and external replications, where new data are collected. External replications are divided into exact, approximate, and conceptual. An exact external replication, like Polio and Gass’s virtual replication, is similar to the original study in every detail which makes them rare, if not impossible, but serve to confirm the original findings. An approximate external replication, also called a partial or systematic replication, contains some minor changes and primarily verifies the original but at the same time provides some evidence of generalisability as far as the minor changes are concerned. Finally, a conceptual external replication addresses the same problem or phenomenon but through a different methodology in an attempt to confirm the robustness of the original study’s claims by providing converging evidence. Given the importance of accuracy over generalisability when using text coverage as a critical value in predicting reading comprehension, an approximate external replication seems to be the most appropriate course to take.


Although Hu and Nation’s (2000) study has attained a respectable status among text coverage researchers, there are several issues which may affect the validity, generalisability, and reliability of the results. These limitations, along with the possible influence they may have had on the results and the text coverage model of reading comprehension, further justify replicating the study. In addition, there is scope to expand the study and potentially improve the predictive power of the text coverage model by addressing
some questions which arose from it. The corrections and improvements can be classified into the following three categories:

- participant heterogeneity
- measurement instruments
- construct validity

2.7.1 Participant heterogeneity

Ideally, models in applied linguistics should be robust with respect to various learner characteristics, so it is understandable that the Hu and Nation (2000) cohort was heterogeneous. This approach to participant selection is common within applied linguistics presumably to increase generalisability of the results. But there are several issues with the selection and measurement, or lack thereof, of certain participant characteristics which may explain some of the variance in L2 reading comprehension. These factors include age, gender, L1, and L2 learning history. Although random assignment to treatment conditions should have negated any differential effects that might arise from these, the possibility can not be ruled out because of the way the text coverage model of reading comprehension is universally applied to all L2 English learners. Randomisation is useful for isolating variables of interest, but it comes at the expense of being able to generalise the effect to distinct groups of learners. Accounting for such potential effects is also a necessary step.

2.7.1.1 Age and L2 acquisition history

Although the participants are likely adults, it is not clear what their ages are. The only information given is that “[m]ost were preparing for undergraduate or post-graduate study in an English speaking university.” (Hu, 1999, p. 35; Hu & Nation, 2000, p. 410) This implies age ranges from 18 to mid- or even late-20s and possibly even older. Although L2 ability is probably not strongly correlated with ages in this range, there are a number of possible scenarios which could be important to take into account which are precluded due to the lack of data. One such scenario is that the younger participants might have a different degree of fluency than older participants even if their vocabulary sizes are roughly the same. This could occur if both older and younger participants studied L2 English for roughly the same continuous period of time but differ in the age at which the studies commenced. Strong evidence exists to support the applicability

---

30 Details on the participants’ vocabulary size as a selection criterion can be found in subsection 3.5.1, page 86.
of Critical Period Hypothesis in L2 learning situations (e.g. DeKeyser, 2000; J. S. Johnson & Newport, 1989), so differences in L2 English acquisition commencement may be an important source of variance which, unfortunately, can not be explained due to the absence of age and language learning history data.

### 2.7.1.2 Gender

In a post-structuralist analysis of reading research methodologies in *Reading Research Quarterly*, Patterson (1995) found that “the term gender or sex is one of the most over-used and undervalued terms in reading research” (p. 290) because descriptive statistics were reported on gender and/or sex in almost half of the studies published during the five year period of Patterson’s study, but that variable was included in the studies’ analyses a mere 2% of the time. Bernhardt (2005) also notes the lack of consideration of gender in reading research, pointing out that most L1 reading models assume that gender may explain differences in literacy.

As an example of a study which did consider gender, Pae (2004) found evidence for a gender differential on performance of two of five categories of reading comprehension question types. Among the 17–19 year old Korean participants, females tended to perform better on items which assessed the ability to recognise the mood, impression, or tone of a passage while males tended to perform better on items which assessed the ability to make logical inferences about a passage. The test conditions were high-stakes and the passages were short, just 80–150 words each, so the results are not directly applicable to the Hu and Nation (2000) participants, but the existence of a difference is intriguing. The majority of the MCQ reading comprehension questions in Hu and Nation (2000), 10 out of 14, assessed inferential propositions, so it is possible that gender may explain some of the reading comprehension variance.

Gender can also affect topic familiarity and recall of text content which is not gender-neutral. Brantmeier (2003) compared the comprehension of two 600 word, gender-oriented texts by American participants enrolled in tertiary-level L2 Spanish courses in a low-stakes reading environment. One text featured an all-female cast of characters describing a frustrated woman visiting a college roommate while the other contained only male characters discussing a boxing match. Brantmeier found no significant difference between genders in overall reading comprehension performance but did detect significant differences in topic familiarity and a significant effect between gender and text content on both MCQ and recall comprehension measures. The text, measures of reading comprehension, and participant ages are generally analogous to the Hu and Nation (2000) participants. The single text used was narrated by a woman caught in
a terrifying practical joke making it roughly analogous to Brantmeier’s (2003) female-oriented text which raises the possibility that male participants could have been at a disadvantage.

There is no account of the gender balance of the participants so if they were primarily female, their performance on a female-oriented text would likely be higher than would be expected over a range of gender-neutral and gender-biased texts. If that is the case, then the model coefficients reported in Hu and Nation (2000) could be exaggerated if generalised to all genders and all texts.

2.7.1.3 L1 backgrounds

Hu and Nation’s (2000) participants were reported to be L1 users of German, Vietnamese, Ni-Vanuatu, Indonesian, Thai, Korean, Chinese, and Japanese. While such a diverse sampling of L1 backgrounds may improve generalisability, there are important differences among them which are probably better investigated than ignored. For example, the first four languages in the list all use Latin-based alphabetic writing systems but Thai and Korean use abugida writing systems, Chinese employs a logosyllabic writing system, and Japanese is a combination of a syllabic, or more accurately a moraic, and logographic writing systems.

It is well-established that writing systems strongly dictate the perceptual strategies used during fluent reading. For example, fluent readers of Thai, which does not use white space to demarcate word boundaries, make use of internalised character frequency statistics to identify word boundaries which they use to determine the appropriate landing site for each fixation (Kasisopa, Reilly, Luksaneeyanawin, & Burnham, 2013). Although L2 learners do eventually adopt native speaker-like eye movements and fixation patterns as proficiency increases (Koda, 1996), there is a very real possibility that participants who lack a shared writing system with English would be handicapped compared with those whose writing system is essentially the same.

A cohort which shares a common L1 are also likely to share a culture. Differences between the culture of the reader and the culture embedded in a narrative text impair reading comprehension. There is evidence, for example, that readers from Middle Eastern cultures can misinterpret some implicit information to misinterpret a text. Yousef (1968) describes a situation in which Middle Eastern students of L2 English were reading Willa Cather’s short story “The Sculptor’s Funeral” in which the cruel mother of a famous sculptor ostentatiously feigns grief at a funeral. In many Middle Eastern cultures, however, the intensity of outward displays of grief is directly proportional to the sin-
cerity of the emotions. As a result of this obvious clash in the cultural interpretations of body language, “[t]hey missed the point completely.” (Yousef, 1968, p. 228)

The potential for misunderstanding is likely further compounded in short fiction and more literary styles in general where the mantra for writers of such genres is show, don’t tell. Implicit details and subtle clues are one of the hallmarks of literary texts. Without explicit explanation, the comprehension of the text relies heavily on a body of knowledge shared between the author and the reader. The text in Hu and Nation (2000), Hill and May’s (1962) “The Escaped Lunatic”, likewise relies on the reader interpreting actions, body language, and responses to events which are similar to the author’s interpretation. Failure to do so will result in a gross misinterpretation of the story. Rather than starting off full of suspense and tension and ending with relief, misinterpretation of the explicit but culturally-laden cues could result in a more neutral interpretation of events or even a categorically different tone.

Yousef’s (1968) study was observational, but a similar phenomenon was confirmed with similar participants in a more rigorous methodology involving texts matched in terms of readability and theme but differing in cultural settings. Nelson (1987) found that recall was significantly higher when texts matched the culture of the adult Egyptian readers even though the readers did not prefer one set of culturally-matched texts over the other. Recall is the essence of delayed reading comprehension tests, so participants in Hu and Nation’s (2000) study whose cultures are more distant from the Western culture embedded in “The Escaped Lunatic” would be disadvantaged, again limiting the generalisability of the results.

2.7.2 Instruments

Although the methodology in Hu and Nation (2000) is sounder than that of correlational studies, there are several issues in the measurement instruments used which, if improved, could increase the reliability and validity of the results.

2.7.2.1 L1 literacy

The most basic interpretation of the hypothesis and short circuit hypotheses requires some skill which is useful in L2 to be blocked from use due to a lack of L2 ability. In terms of reading, this must be some generalised reading skill which becomes available as text coverage increases. Yet this skill is never measured directly in Hu and Nation (2000). Therefore we are left to assume that the L1 reading skill which would be blocked below the threshold is somehow invariant across participants. Randomisation should
limit the effect of differences within the cohort, but there is still no assurance that any of them have the required skill in the first place since it is not measured. We can infer somehow that the participants must have some reasonable level of L1 reading ability by the fact that they were post graduate students attending a university in an English-speaking country, but that is simply not enough when the entire theory rests on that skill becoming available once the threshold has been surmounted. Vocabulary size was measured, and used as a selection criterion for inclusion in the study. Some participants were even excluded from the study based on their sub-par performance on the test, but a simple measure of L1 literacy was not included at all. We would expect that, at minimum, a measure would be taken of L1 literacy at least to determine who should be included in the study. Additionally, if more sophisticated statistical analyses are to be used, such as multiple linear regression which can account for other explanatory variables, then L1 literacy must be measured somehow.

2.7.2.2 Reading comprehension

Despite the availability of standardised measures of reading comprehension which are widely used in the fields of psychology, psycholinguistics, and, to an extent, in linguistics, standard practice in applied linguistics appears to favour self-created measures. This is not wholly unjustified since the nature of applied linguistics implies the investigation of processes and phenomena which are robust to materials and procedures which are not standardised and are often created ad hoc by language practitioners and materials creators on an as-needed basis.

However, comparisons can be made between the nature of standardised reading comprehension tests and the test employed in Hu and Nation (2000). The Scholastic Aptitude Test (SAT), for example, is a standardised test used by higher education institutions in the United States of America as part of the admissions process. One component of this test is the Critical Reading component which comprises four reading passages of varying lengths and a total of forty questions. Not all of the questions measure reading comprehension; some of them measure vocabulary knowledge in context as well as literal comprehension and extended reasoning. Another well-known standardised test, the Nelson-Denny Reading Test, is commonly used to identify individuals in need of remedial reading instruction at the secondary and tertiary levels. The reading comprehension component of the test is often used in research because it comes in two parallel forms, G and H, which consist of seven passages and 38 questions.

Other tests are similar and show a trend of around 40 questions to accurately measure reading comprehension. The significance of this property can be inferred from research

which attempted to use a shortened form of a reading comprehension measure only to find that doing so negatively affected the power of the experiment design. For example, Daneman and Hannon (2001) measured the effect of four different reading test strategies but apparently only had access to two complete SAT tests. Daneman and Hannon assigned half of each form to each condition and, after failing to find the hypothesised statistical difference, concluded that using half of each test “may not have provided sufficient data for us to pick up reliable effects.” (p. 214). This conclusion is supported by a second study where the entire form was used and the difference in strategy that the results from the first study suggested was detected.

Although using a single passage and a reasonable number of comprehension questions for that passage may be ecologically valid, it increases the probability that the results obtained could have been a fluke. In Hu and Nation (2000), the MCQ reading comprehension measure comprised only 14 questions and seemed to show a clear ceiling effect. The written cued recall measure, in contrast, comprised a more reliable 27 questions which did not appear to exhibit a ceiling effect. It is possible that the parameters of the MCQ measure might be different if additional questions are added to avoid the ceiling effect and bring the measure more in line with accepted standardised measures of reading comprehension such as the SAT and the Nelson-Denny Reading Test.

2.7.2.3 Consistency of pseudowords

Texts with known text coverage figures were created by replacing infrequent words with pseudowords. Failure to maintain consistency in the pseudowords, however, would result in the participants being unable to draw upon skills used in guessing the meaning of unknown words from context, thus making the task more difficult than it actually should be. An example of the pseudoword inconsistencies in the original study’s texts can be seen in Table 2.1 (page 47).

Additionally, some of the pseudowords in the original treatment texts could easily be confused with real words. For example, the pseudoword *pertruding* is orthographic-
ally and phonetically similar to the real word *protruding*. If it was read as such by the participants, the confusion brought on by the incongruous usage in the given context could have made the text more difficult that it should have been.

### 2.7.3 Construct validity

Hu and Nation conclude their study supported the claim made by Hirsh and Nation (1992) that 98% text coverage is needed “to be able to read for pleasure.” (p. 422) The text coverage figure itself is substantiated by data, but the reading conditions in which that figure applies must be drawn from the conditions present during the data collection. The procedure detailed in Hu and Nation (2000) is too brief to envision the conditions, but an appendix in Hu (1999) contains the instructions given to the participants:

We want to find out about unknown vocabulary and your understanding of the passage. Each of you will be given a different reading passage which will contain some unknown words. Do not worry too much about those unknown words while reading.

Read the passage. You can read it as many times as you like. Afterwards, you will be asked to answer some questions and write down as much information as you could remember from the text without time limit, but you cannot look back at the text. Then the second comprehension measure you will be asked to do is the multiple choice questions. The results of those tests will only be used for our research. (Hu, 1999, p. 146)

Hu and Nation distinguish between three types of reading which might be done by L2 English readers: intensive with support, extensive for language growth, and extensive for fluency development. Of the three, the instructions given to the participants clearly match the *extensive for fluency development* type since there was no support available to the participants and time was not constrained which would induce the speed needed to build fluency. But is the list of possible reading scenarios complete?

Carver (2002) asserts that reading can be divided into five processes, summarised in Table 2.2 (page 49), each of which requires the use of distinct mental processes. The difference between Carver’s categories and Hu and Nation is that the former is concerned with what is involved in performing the task whereas the latter is focused on the purpose.

The term *reading for pleasure* explicitly deals with purpose but the processes involved in achieving that purpose can not be ignored. To demonstrate with Carver’s five processes as an example, few would concede that memorising takes place when reading for
Table 2.2: Five Basic Reading Processes according to (Carver, 2002).

<table>
<thead>
<tr>
<th>Process</th>
<th>Culminating Component</th>
<th>Typical Rate for College Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanning</td>
<td>Lexical accessing</td>
<td>600 WPM</td>
</tr>
<tr>
<td>Skimming</td>
<td>Semantic encoding</td>
<td>450 WPM</td>
</tr>
<tr>
<td>Rauding</td>
<td>Sentence integrating</td>
<td>300 WPM</td>
</tr>
<tr>
<td>Learning</td>
<td>Idea-remembering</td>
<td>200 WPM</td>
</tr>
<tr>
<td>Memorising</td>
<td>Fact-rehearsing</td>
<td>138 WPM</td>
</tr>
</tbody>
</table>

pleasure. Clearly reading for pleasure must be associated with rauding (Carver’s term for normal reading), learning, or some point in between. Is it valid, then, to assume that the rauding and/or learning processes were used by the participants they know in advance that they would be tested on the content of the text? It is plausible that in this case, the participants engaged in primarily learning if not some memorising in preparation for the reading comprehension tests. If that is so, then the validity of the claim that the participants were reading for pleasure must be questioned. Although this does not in any way diminish the validity of the relationship between text coverage and reading comprehension, it does cast doubt on the circumstances under which this relationship applies.

2.8 Conclusion

The history of the relationship between vocabulary and reading is long and complex. Out of necessity, only a very narrow set of studies have been reviewed here even though many other topics and studies are relevant and would benefit greatly this chapter had they been reviewed. The operationalisation of the term *reading comprehension*, for example, has undergone a greater evolution over the course of the past century than *text coverage*. It is a fundamental aspect of the reading process, yet there is not yet any universally agreed definition nor operationalisation (Garman, 1990).

Nevertheless, this chapter has traced the evolution of the modern conception of text coverage from its humble roots in education reform, word frequency lists, and the application of science to the teaching of reading. Related, and often conflated, goals such as optimising the rate of vocabulary acquisition and ranking the reading difficulty of texts has also been explored and, where needed, artificially excised from the focus of this thesis. Studies which build directly on the text coverage model of reading compre-
hension, such as those recommending vocabulary sizes needed comprehending various texts, were also discussed.

Ultimately, the veracity of the text coverage model rests primarily on the results obtained in Hu and Nation (2000). This situation strongly justifies a complete replication of the core elements of the study as well as expanding the range of possible predictor variables. If the replication confirms the initial results and other variables can be added to text coverage model to increase its predictive power, then the field of applied linguistics will benefit from an increased understanding of the most important factors in successful L2 English reading comprehension.
Table 2.3: Summary of indicative research related to the text coverage model of reading comprehension.

<table>
<thead>
<tr>
<th>Operationalisation</th>
<th>Vocabulary Burden / Text Coverage</th>
<th>Reading Comprehension</th>
<th>Language Proficiency</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. L. Johnson (1977)</td>
<td>participants: 18; texts: 2; text length: unknown; variance explained: unknown</td>
<td>• Inferred vocabulary burden by ranking texts proportion of “difficult” words • “Difficult” operationalised as infrequent when compared to a corpus-based frequency list, adjusted by frequency counts from textbooks</td>
<td>• Read two texts differing in difficulty • Asked participants which of the two was − read first − enjoyed more − more difficult</td>
<td>• Not measured</td>
</tr>
<tr>
<td>Operationalisation</td>
<td>Vocabulary Burden / Text Coverage</td>
<td>Reading Comprehension</td>
<td>Language Proficiency</td>
<td>Conclusions</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------------</td>
<td>----------------------------------</td>
<td>-----------------------</td>
<td>----------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Holley (1973); participants: 246; texts: 1; text length(s): 636; variance explained: unknown | • Grammatical and lexical simplifications  
• 50 words with low frequency synonyms selected  
• Created versions of simplified text with 5, 10, 20, 25, 30, 40, or 50 low frequency synonyms substituted | • 10 item MCQ comprehension test  
• primarily explicit propositions, two implicit  
• "Although the reliability of the test (alpha coefficient of internal consistency = .78) is fairly high, the validity and sensitivity of the test are probably not as high as desirable. Further research upon the relation between comprehension and vocabulary density seems necessary." (pp. 343–346) | • Not measured  
• Participants had all taken the same courses previously, so assumed to be of similar ability | • No significant effect on comprehension  
• Data do not support a maximum or optimal rate of unknown vocabulary for acquisition purposes |
<table>
<thead>
<tr>
<th>Operationalisation</th>
<th>Vocabulary Burden</th>
<th>Text Coverage</th>
<th>Reading Comprehension</th>
<th>Language Proficiency</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laufer (1989)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>participants: 100; texts: 1; text length(s): unknown; variance explained: unknown</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-reported by underlining unknown words and verified through translation task of most frequent words from the text. Assumed that all words not tested in the translation were known. Post-hoc groups based on reported knowledge of ≥5% or &lt;5% of the words in the text.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥55% on institutional reading comprehension test which tested comprehension of general ideas to elicit and interpret supporting information. Can be inferred from text coverage. Assumed all participants similar in ability.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not measured directly. Can be inferred from text coverage. Assumed all participants similar in ability.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥95% group had a significantly higher number of subjects attaining passing marks on reading comprehension test than &lt;95% group.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not measured directly. Can be inferred from text coverage. Assumed all participants similar in ability.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Laufer and Sim (1985) |                  |               |                      |                      |             |
| participants: >12; texts: 3; text length(s): unknown; variance explained: unknown |
| Oral and written questions distinguishing between main and peripheral ideas; distinguishing between explicit and implicit material; recognizing the author's intent; and interpreting supporting information. |
| Reading comprehension section of the Cambridge First Certificate in English (FCE). |
| Author-made Reading Strategies Exam. |
| 65—70% on the FCE represents a threshold for reading comprehension competence of academic texts. |

Table 2.3: Summary of indicative research related to the text coverage model of reading comprehension (continued)
Table 2.3: Summary of indicative research related to the text coverage model of reading comprehension. (continued)

<table>
<thead>
<tr>
<th>Operationalisation</th>
<th>Vocabulary Burden / Text Coverage</th>
<th>Reading Comprehension</th>
<th>Language Proficiency</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carver (1994)</td>
<td>• Self-reported unknown words by underlining</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Not measured but cited Carver (1990) which found a negative linear relationship between relative difficulty and ability of a passage to be comprehended when engaged in “rauding” (i.e. normal reading p. 414)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Accuracy Levels Test (vocabulary test, $n = 100$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Relatively easy texts ($\approx 100%$ text coverage) contain $\approx 0%$ unknown words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Texts matched to reader ability ($\approx 90%$ text coverage) contain $\approx 1%$ unknown words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Relatively difficult texts ($\leq 98%$ text coverage) contain $\geq 2%$ unknown words</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3: Summary of indicative research related to the text coverage model of reading comprehension. (continued)

<table>
<thead>
<tr>
<th>Operationalisation</th>
<th>Vocabulary Burden / Text Coverage</th>
<th>Reading Comprehension</th>
<th>Language Proficiency</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu and Nation (2000)</td>
<td>participants: 66; texts: 1; text length: 673; variance explained: 49–62%</td>
<td>• Controlled through carefully designed texts interspersed with pseudowords</td>
<td>• MCQ ($n = 14$) targeting 9 implicit propositions and 5 explicit propositions; $\geq 12$ considered adequate</td>
<td>• VLT performance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Cued written recall ($n = 124$) targeting mostly explicit propositions; $\geq 70$ considered adequate</td>
<td>• $\geq 14$ on 2,000 word level, 18 possible</td>
<td>• $\geq 10$ on 3,000 word level, 18 possible</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\bar{x} = 63.49[37, 89]$; SD = 11.80, 90 possible</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3: Summary of indicative research related to the text coverage model of reading comprehension. (continued)

<table>
<thead>
<tr>
<th>Operationalisation</th>
<th>Vocabulary Burden / Text Coverage</th>
<th>Reading Comprehension</th>
<th>Language Proficiency</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Web-based lexical profiler (<a href="http://lextutor.ca">http://lextutor.ca</a>) used to calculate lexical profile of three older versions of the reading comprehension test comprising 19,037 running words</td>
<td>• Reading section of the Psychometric University Entrance Test designed and marked by the National Institute for Testing and Evaluation (Israel) comprising 60 timed multiple choice questions</td>
<td>• Items target vocabulary knowledge, syntactic knowledge, both implicit and explicit global textual propositions, global and local textual purpose, and making inferences</td>
<td>• 2,000, 3,000, and 5,000 sections of the VLT</td>
<td>• Two text coverage thresholds: 98% for optimal comprehension representing knowledge of ≈ 8000 words and 95% for minimal comprehension representing knowledge of ≈ 4000 – 5000 words</td>
</tr>
<tr>
<td>- Proper nouns recognised by lexical profiler considered as 1k level words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 95% text coverage (actually 94.91%, but rounded up) equates to 4,000 word families, 97.5% to 98.4% text coverage equates to 7,000 – 8,000 word families</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Laufer and Ravenhorst-Kalovski (2010)
participants: 745; texts: unknown; text length(s): unknown; variance explained: 64%
Table 2.3: Summary of indicative research related to the text coverage model of reading comprehension. (continued)

<table>
<thead>
<tr>
<th>Operationalisation</th>
<th>Vocabulary Burden / Text Coverage</th>
<th>Reading Comprehension</th>
<th>Language Proficiency</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prichard and Matsumoto (2011)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>participants: 103; texts: 1; text length(s): 650; variance explained: 8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Directly measured knowledge of 71 word types using a translation test</td>
<td>- MCQ (n = 8) targeting main ideas, explicit propositions, and author’s stance</td>
<td>- Not measured, but reported as &quot;lower-intermediate to intermediate&quot; (p. 213)</td>
<td>- Reading comprehension significantly increases at 92–93% text coverage</td>
<td></td>
</tr>
<tr>
<td>- Assumed all untested word types were known</td>
<td>- Questions in participants’ L1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Yoon (2011)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>participants: 41; texts: 20; text length(s): 64–86; variance explained: 24%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Controlled through carefully designed texts interspersed with pseudowords</td>
<td>- Each participant read four texts at each text coverage level</td>
<td>- Not measured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Five levels: 100%, 95%, 90%, 85%, and 80%</td>
<td>- 2 questions per text</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Questions focused on main ideas, explicit details, and implicit details</td>
<td>- Questions focused on main ideas, explicit details, and implicit details</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

95% text coverage enables most L2 Korean readers to achieve unassisted adequate comprehension.
Table 2.3: Summary of indicative research related to the text coverage model of reading comprehension. (continued)

<table>
<thead>
<tr>
<th>Operationalisation</th>
<th>Vocabulary Burden / Text Coverage</th>
<th>Reading Comprehension</th>
<th>Language Proficiency</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. Schmitt et al. (2011)</td>
<td>participants: 661; texts: 2; text length(s): 582–757; variance explained: 17%</td>
<td>• Directly measured knowledge of &gt; 50% of word types in texts using a Yes/No checklist</td>
<td>• Text available</td>
<td>• Not measured directly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Proportion of Yes responses for each frequency level on checklist</td>
<td>• MCQ ( (n = 14) ) targeting main idea, explicit and implicit propositions, author’s stance, and genre analysis</td>
<td>• Can be inferred from text coverage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>used to extrapolate knowledge of types in text belonging to equivalent frequency levels</td>
<td>• Graphical organisers ( (n = 16) ) targeting explicit propositions</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

External replication and expansion of
Hu and Nation (2000)

This study described in this chapter builds upon Hu and Nation (2000) in four distinct ways. First, by using the original text and MCQ comprehension measures, this will verify whether regression coefficients and explained variance were the result of a real phenomenon or simply a fluke. Second, by decreasing the variability in the participants, it will investigate the generalisability of the results to other, more homogeneous populations. Third, by correcting some minor errors in the original instruments and design, the sensitivity of the measurements should be improved. Finally, by adding other measures thought to account for variance in reading comprehension, such as L1 literacy (attitudes and practice) and L2 proficiency (vocabulary size), this study will attempt to increase the explainable variance in reading comprehension which would lead to a more complete, and useful, model.

3.1 Research questions

The first two research questions in this study are, by necessity, close equivalents of those in the published study, Hu and Nation (2000), and the unpublished Master’s thesis, Hu (1999). The reason why the research questions here are not exactly the same is because the two source studies do not share exactly the same questions. So one needs to be chosen over the other or hybrid questions need to be created. The unpublished study contained the following three research questions:

1. Will different densities of unknown words result in differences in comprehension? In particular, as the number of unknown words increases, will comprehension decline?

2. Is 95% coverage the threshold level for successful reading comprehension of a fiction text?
3. What is the most appropriate density of unknown words for readers to achieve pleasurable reading as well as vocabulary learning? (pp. 33–34)

Research Question Three was not included in the published study. One obvious problem is that it assumes both pleasurable reading and vocabulary learning can take place at the same time. If that is not the case, then the question is essentially a double-barrelled question which is a form to be avoided when constructing questionnaires (Dörnyei, 2007). Even if it is possible for the two goals to occur concomitantly, the methodology used to answer the question did not rise to the level of robustness (see Hu, 1999, p. 90) required for publication in a peer-reviewed journal.

That was not the only change, though. The research questions which appeared in Hu and Nation (2000) are:

1. Will different densities of unknown words result in differences in comprehension? In particular, as the number of unknown words increases, will comprehension decline?

2. Is there a vocabulary coverage level which acts as a threshold between adequate and inadequate comprehension a fiction text? (p. 410)

Note that Research Question Two has subtly changed from testing a claim from previous research to a more generalised search without a preconceived figure. As this study is a replication, it would be wise to revert back to the question of verifying the accuracy of previous results.

Since it has been established that there is a predictable, quantified relationship between unknown word density, i.e. text coverage, and reading comprehension, the first research question in this study can be rewritten as a verification of Hu and Nation’s (2000) results using the ideas in that study’s first research question:

**Research Question One:** Are the coefficients of reading comprehension regressed on text coverage reported in Hu and Nation (2000) the same for data from a different population?

This question is thematically the same as the Hu and Nation’s Research Question One, but it incorporates the new knowledge gained from the previous studies. If the regression coefficients describing the relationship are the same, or at least within the confidence intervals, then the claim will be supported and confidence in the robustness of the relationship will be increased.
3.1. RESEARCH QUESTIONS

For the second research question, the original studies established that 98% represents a text coverage level which provides the majority of readers with the opportunity to achieve adequate reading comprehension. This study intends to verify that claim, so the second research question can be rewritten in the spirit of Hu’s (1999) second research question:

**Research Question Two:** Does the 98% text coverage figure reported in Hu and Nation (2000) represent a level at which the majority of readers from a different population can achieve adequate comprehension?

As with Research Question One, if Research Question Two is answered in the affirmative, the 98% coverage figure can be seen as more reliable which strengthens the foundation of research which is based on the figure (e.g. I. S. P. Nation, 2006).

Finally, Hu (1999) found no significant relationship between estimates of vocabulary size derived from the VLT and the two measures of reading comprehension, but when the relationship was examined for each individual text coverage treatment, there was a fairly strong significant correlation of $r = .61$ between the MCQ reading comprehension measure and the vocabulary size measurements for the 100% text coverage treatment. These correlations are reproduced in Table 3.1 (page 62). The magnitude of the correlation diminished with text coverage and was not significant for any of them. These results seem to be conflicting. On the one hand, the lack of correlation is dismissed as “not surprisingly [sic] because vocabulary size and coverage of known words are confounded” (Hu, 1999, p. 88), yet the decreasing correlation with text coverage suggests an interaction between the two. Participants were randomly assigned to treatments and text coverage was determined by the inclusion of pseudowords, so, unlike observational studies, vocabulary size and text coverage should not be confounded. With a more sensitive and accurate measure of vocabulary size, and a statistical model which accounts for both vocabulary size and text coverage, the relationship between all three may start to become more apparent. At the heart of the threshold hypotheses which originally motivated this line of research, other factors such as L1 literacy are also thought to be involved in L2 reading. Therefore, the third research question for this study is:

**Research Question Three:** Does the addition of vocabulary size or measures of literacy in L1 or L2 as explanatory variables increase the amount of meaningfully explained variance in reading comprehension?

---

1 Further details on observational text coverage studies which confound vocabulary size and text coverage can be found in section 2.5 (page 34).
Table 3.1: Vocabulary size and reading comprehension correlations by text coverage treatment reported in Hu (1999, p. 88). * indicates $p < .0001$.

<table>
<thead>
<tr>
<th>Coverage (%)</th>
<th>MCQ</th>
<th>Cued recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>.61 *</td>
<td>.24</td>
</tr>
<tr>
<td>95%</td>
<td>.46</td>
<td>.24</td>
</tr>
<tr>
<td>90%</td>
<td>.36</td>
<td>-.39</td>
</tr>
<tr>
<td>80%</td>
<td>.17</td>
<td>.04</td>
</tr>
</tbody>
</table>

3.2 Participants

The participants in Hu and Nation (2000) were recruited from a variety of L1 backgrounds and were studying English in New Zealand when the data were collected.² It terms of generalisability, this sample may be seen as fairly representative of advanced L2 English learners studying in English-speaking countries at the tertiary level.

In an effort to ensure reasonable homogeneity of L1 background, cultural schemata, background knowledge, topic familiarity, and other factors which may play a role in reading comprehension, 138 students from four intact compulsory first year English courses at a university in northern Thailand were invited to participate. The selection of this group represents a compromise between exactly matching that in Hu and Nation to achieve an exact replication, slightly modifying some characteristics in order to achieve an approximate replication, and refining the methodology to meet best practises as set forth advocated by Bernhardt (2011).

A few key differences should be noted. The participants in the present study were studying at the tertiary level, but at a much earlier stage than that of Hu and Nation and therefore slightly younger as well. The L1 and cultural backgrounds of the participants were mostly homogeneous, unlike the mixed group in Hu and Nation. Perhaps most importantly, the participants were studying English in their home country where English is not widely spoken although it does represent a strong unofficial second language. This contrasts sharply with the participants in Hu and Nation who were all sufficiently skilled and motivated enough to study abroad where English is the medium of instruction. The following subsections examine five key participant characteristics in detail: age, gender, linguistic landscape, L2 English learning history, and literacy.

²Further details on the problem of homogeneity can be found in subsubsection 2.7.1.3, page 44.
3.2. PARTICIPANTS

3.2.1 Class and age

Four intact classes were invited to participate in this study. Although the classes had just commenced a few weeks prior to data collection, the instructor was asked to subjectively rank the classes in terms of perceived overall L2 English ability relative to the other classes in this study. Labels were assigned to each class to reflect the ranking reported by the instructor. A brief description of the classes, subjective assessments of general English ability given by the instructor, and the labels used in this study to refer to them are as follows:

Class A: Second-year students majoring in medicine; strong English ability

Class B: Second-year students majoring in engineering, science, agriculture, and agro-industry; average to weak English ability

Class C: First-year students from a variety of faculties; lower or weak English ability

Class D: First-year students from a variety of faculties; lower or weak English ability

Because the courses were compulsory in the first and second years of study, the ages of the participants are generally equivalent which can be seen in Table 3.2 (page 64).

3.2.2 Gender

At the end of the reading attitudes survey (see subsection 3.3.5, page 79), participants were asked to register their เพศ (pʰêːt̚, sex or gender). Mindful of Patterson’s (1995) admonishment of the reading research community for treating gender as a bodily property, participants were not given options from which to choose; they were free to write any term describing the gender community with which they most self-identified in the space provided and in so doing, escape the “binary structuralism” (Patterson, 1995, p. 294) imposed in much reading research. It is not uncommon in Thailand to consider at least three gender categories when collecting demographic information. These categories generally include ผู้ชาย (pʰûː ʨaːj/, male), ผู้หญิง (pʰûː yǐŋ/, female), and ผู้หญิงประเภทสอง (pʰûː yǐŋ pràʔ.pʰêːt̚sɔ̌ːŋ/, second type of female).

Despite offering the freedom to report any gender identity and the author’s observation that some participants had the outward appearance of belonging to a third gender, participants only reported ผู้หญิง (pʰûː yǐŋ/, female) or ผู้ชาย (pʰûː ʨaːj/, male) when responding to the question of gender. Nevertheless, this self-classifying approach should avoid the problem of assigning gender labels by using a “sight test” (Patterson, 1995, p. 294).
Table 3.2: Distribution of text coverage treatment and class by participant gender and age.

<table>
<thead>
<tr>
<th>coverage</th>
<th>ผู้หญิง (female)</th>
<th>ผู้ชาย (male)</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18 19 20 n</td>
<td>18 19 20 n</td>
<td>18 19 20 n</td>
</tr>
<tr>
<td>86%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>6 2 0 8</td>
<td>0 4 0 4</td>
<td>6 6 0 12</td>
</tr>
<tr>
<td>B</td>
<td>0 2 3 5</td>
<td>0 1 3 4</td>
<td>0 3 6 9</td>
</tr>
<tr>
<td>C</td>
<td>3 4 0 7</td>
<td>4 0 0 4</td>
<td>7 4 0 11</td>
</tr>
<tr>
<td>D</td>
<td>3 1 0 4</td>
<td>5 3 0 8</td>
<td>8 4 0 12</td>
</tr>
<tr>
<td>n</td>
<td>12 9 3 24</td>
<td>9 8 3 20</td>
<td>21 17 6 44</td>
</tr>
<tr>
<td>94%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>3 3 0 6</td>
<td>0 5 0 5</td>
<td>3 8 0 11</td>
</tr>
<tr>
<td>B</td>
<td>0 3 4 7</td>
<td>0 2 1 3</td>
<td>0 5 5 10</td>
</tr>
<tr>
<td>C</td>
<td>6 1 0 7</td>
<td>3 0 0 3</td>
<td>9 1 0 10</td>
</tr>
<tr>
<td>D</td>
<td>3 5 0 8</td>
<td>3 1 1 5</td>
<td>6 6 1 13</td>
</tr>
<tr>
<td>n</td>
<td>12 12 4 28</td>
<td>6 8 2 16</td>
<td>18 20 6 44</td>
</tr>
<tr>
<td>98%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5 5 0 10</td>
<td>1 2 0 3</td>
<td>6 7 0 13</td>
</tr>
<tr>
<td>B</td>
<td>0 6 2 8</td>
<td>0 1 0 1</td>
<td>0 7 2 9</td>
</tr>
<tr>
<td>C</td>
<td>2 3 0 5</td>
<td>1 5 0 6</td>
<td>3 8 0 11</td>
</tr>
<tr>
<td>D</td>
<td>3 3 0 6</td>
<td>3 2 0 5</td>
<td>6 5 0 11</td>
</tr>
<tr>
<td>n</td>
<td>10 17 2 29</td>
<td>5 10 0 15</td>
<td>15 27 2 44</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>14 10 0 24</td>
<td>1 11 0 12</td>
<td>15 21 0 36</td>
</tr>
<tr>
<td>B</td>
<td>0 11 9 20</td>
<td>0 4 4 8</td>
<td>0 15 13 28</td>
</tr>
<tr>
<td>C</td>
<td>11 8 0 19</td>
<td>8 5 0 13</td>
<td>19 13 0 32</td>
</tr>
<tr>
<td>D</td>
<td>9 9 0 18</td>
<td>11 6 1 18</td>
<td>20 15 1 36</td>
</tr>
<tr>
<td>n</td>
<td>34 38 9 81</td>
<td>20 26 5 51</td>
<td>54 64 14 132</td>
</tr>
</tbody>
</table>
3.2. PARTICIPANTS

p. 294). The distribution of participant gender delineated by class\(^3\) and treatment condition\(^4\) are detailed in Table 3.2 (page 64). Note that the majority of participants are female.

It should also be noted that, from the author’s personal subjective experience teaching L2 English at the tertiary level in Thailand for two years, students who appeared to identify more with female-ish genders seem to be more adept at L2 acquisition, both in taught language environments and in their use of L2s in more quotidian situations, than those who identify with more male-ish genders. Based on that first-hand experience, it is compelling to consider gender as an independent variable which may add some explanatory power to the text coverage model of reading comprehension.

Ironically, there is some evidence that gender might not play much of a role in reading comprehension for the participants in this study. Phakiti (2003) compared performance on a high-stakes multiple choice cloze and reading comprehension tests along with a survey of cognitive and metacognitive strategies employed while taking the test. The participants were Thai, had experience studying English for about eight years, ranged in age from 17 to 21, and were all taking a compulsory English course. These demographic characteristics compare favourably with the participants in this study. No significant difference was found between genders on reading comprehension test performance nor cognitive strategy use, but males were found to use more metacognitive strategies (Phakiti, 2003). In attempting to explain the results which contradicted most previous research, Phakiti points to the same homogeneity of features which make this demographic group ideal for the type of reading research advocated by Bernhardt (2011):

“[T]hey were all from the same age group, had a relatively small range of English proficiency levels, and shared the same cultural and educational context [...] [T]hese learners had a nonalphabetic L1 coding system.” (Phakiti, 2003, p. 669)

While the results are surprising, perhaps the most interesting conclusion to draw is that the participants Phakiti describes, which closely mirror this study’s participants, are largely homogeneous especially when compared to L2 English reading studies carried out in English speaking countries consisting of participants from a clearly self-selected group. These unique characteristics should reduce any noise inherent in more heterogeneous groups thereby creating the potential to more sensitively investigate the relationship between text coverage and reading comprehension.

\(^3\)Further details on the classes participants were enrolled in can be found in subsection 3.2.1, page 63.

\(^4\)Further details on treatment conditions can be found in subsection 3.3.1, page 69.
3.2.3 Linguistic landscape

One of the difficulties in comparing research on L2 reading comprehension is the failure to clearly delineate the L1 backgrounds of the participants (Bernhardt, 2011), so to avoid that pitfall, a brief overview of the language environment of the participants is warranted. Inhabitants of northern Thailand, where the data were collected, primarily speak a dialect of Thai known locally as คำเมือง (/kʰam mɯaŋ/, Town Language) or formally as ภาษาถิ่นพายัพ (/pʰaːsǎː tʰìn pʰaː.jáp/, Northern Thai Language). It is “significantly and sometimes systematically different” (Smalley, 1994, p. 73) to the syntax, lexis, and phonology of ไทยกลาง (/tʰaj klaːŋ/, Standard Thai). Traditionally, the dialect was written with a unique orthography, the Tai Tham script, but Standard Thai orthography, the Thai script, has largely replaced it. As universities in Thailand have an admissions quota system which ensures that students from various provinces throughout the country are granted admission, all tertiary education students become proficient in Standard Thai as it is the medium of instruction throughout the country. Standard Thai is thus the de facto lingua franca. Although many young Thai students who speak minority languages drop out of the education system (Smalley, 1994), tertiary students can be considered as either L1 users of Standard Thai or, at minimum, proficient users.

English is also part of the linguistic landscape of Thailand; it can even be considered the country’s unofficial second language. It is widely understood to some degree among more educated citizens and is a compulsory subject in primary education. This fact can be seen in the homogeneity of self-reported time spent studying English shown in Table 3.3 (page 68). There is at least one daily English newspaper, Post Today, and English edutainment programmes or segments in broadcast and print media are common.

English orthography is also pervasive although much less so than Standard Thai. Street signs, for example, are commonly written in both Thai and English orthographies. Two standards for transliterating Thai into Latin characters, ISO 11940-2 and the nearly identical Royal Thai General System of Transcription, ensure that most Thais can write their names in the Latin alphabet. The opportunities to encounter English orthography in every day life makes it safe to assume that the participants in this study have reasonably fluent access to English orthography.

Many English textbooks are often used at the university level, especially in technical fields such as computer science and medicine, some of the primary fields of study for a large portion of the participants in this study. When a Thai equivalent or translasi

---

5 This is the spelling, using Standard Thai orthography, which most closely matches the regional pronunciation. It is spelled คำเมือง (/kʰam mɯaŋ/, City Language) in Standard Thai.
3.2. PARTICIPANTS

information is available, however, the Thai version is usually preferred. Many schools in the primary education sector, and increasingly in the secondary and tertiary education sectors, offer English as the medium of instruction for a significant portion, or even the entirety, of the curriculum. Post graduate degrees taught wholly in English are not uncommon and many top-tier universities administer locally produced general English exams as a routine part of their admission process. For example, มหาวิทยาลัยธรรมศาสตร์ (má.hǎː.wíʔ.tʰá.jaː.laj tʰám.má.sàːt̚, Thammasat University), considered one of the top universities in Thailand, requires all undergraduate and graduate students to sit the Thammasat University General English Test (TU-GET).

Despite the prominence of Standard Thai and English, Thailand, like many countries, is actually a very linguistically diverse area containing a large number of minority languages. While the rather uncompromising language policy of imposing Standard Thai on all language groups throughout the country may be seen as detrimental to those communities (Smalley, 1994), for the purposes of this study, it is advantageous. One of the fundamental assertions of Bernhardt (2011) is that measuring L2 reading comprehension should be done in the readers’ L1 to avoid the potential of confounding lack of reading comprehension with difficulty in comprehending the instrument itself. In the present study, the reading comprehension measures were translated to Standard Thai with the assumption that all participants are fluent in it, or at least much more fluent in Standard Thai than in English.

3.2.4 L2 English acquisition history

Participants were asked to estimate the number of years they had been learning English and the number of months they had spent abroad in an environment where English is considered to be the dominant language. Descriptive statistics of the participants’ responses can be found in Table 3.3 (page 68). It is not surprising that the years of learning English figure exhibits a very small amount of variance due to the compulsory courses in English. There is, however, at least one participant who claimed to have studied English for only a single year, which is very unlikely. It is possible that this participant understood the question to refer to the time spent studying English in a tertiary setting.

The responses for the time spent abroad indicate that misunderstanding questions did occur on some occasions even though they were written in Standard Thai and intended to be straightforward. One participant claimed to have spent 216 months abroad.

F6Further details on the participants’ linguistic landscape can be found in subsection 3.2.3, page 66.
### Table 3.3: Participant L2 English learning history

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>mean ($\bar{x}$)</th>
<th>median</th>
<th>min</th>
<th>max</th>
<th>sd (σ)</th>
<th>skewness</th>
<th>kertosis</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years learning L2</td>
<td>125</td>
<td>13.08</td>
<td>13</td>
<td>1</td>
<td>19</td>
<td>2.93</td>
<td>-0.76</td>
<td>1.66</td>
<td>0.00</td>
</tr>
<tr>
<td>Months in L2 area</td>
<td>112</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>1.37</td>
<td>5.05</td>
<td>27.30</td>
<td>0.00</td>
</tr>
</tbody>
</table>
3.3. MATERIALS

which is equivalent to 18 years. This is clearly in error and has been treated as missing. Based on the remaining responses, it is clear that the time spent abroad in English-dominant environments is essentially zero.

3.2.5 Literacy

Reading time per week in both Thai and English are summarised in Table 3.4 (page 70). The figures were derived from self-reported estimates of reading time for different types of text. Participants were asked to reflect on the month-long semester break which had concluded just a few weeks prior to the data collection in this study and estimate the hours per week they spent reading each type of text. Thai university students are required to read large amounts of text during the school year which they may not choose to read on their own, so by focusing on a period of time when reading would likely have been due to their own motivation, it was hoped that the estimates would more closely reflect the participants’ relative literacy capacities.

Post hoc estimates of personal behaviour, unfortunately, are notoriously inaccurate and this phenomenon can be seen in some participants who reported average reading times which exceeded the actual number of hours in a week (168). Rather than simply take these retrospective estimates at face value, they can be considered as a subjective magnitude estimation and therefore used as a rough rank-order estimate of relative literacy as demonstrated by the act of self-initiated reading.

3.3 Materials

This section details changes to the materials used in this study compared with those used in Hu (1999) and Hu and Nation (2000). Where deficiencies were found, they were corrected in a manner which, it is hoped, increased the validity, sensitivity, and accuracy of the study without altering the essential characteristics to such a degree that the present study could not be considered a replication. Supplemental instruments were also used in order to better conform to the more nuanced study of reading comprehension advocated by Bernhardt (2011).

3.3.1 Experimental text

The original narrative text, a short story entitled The Escaped Lunatic, was used. It is a suspenseful story narrated by a woman who is home alone and receives a prank call which negatively influences her interpretation of several serendipitous events which follow in rapid succession. The text was published as part of a collection of thirty short pieces designed to assist students in preparing for the Cambridge Certificate of

7Further details on deficiencies in Hu and Nation (2000) can be found in section 2.7, page 41.
<table>
<thead>
<tr>
<th>Reading</th>
<th>( n )</th>
<th>mean (( \bar{x} ))</th>
<th>median</th>
<th>min</th>
<th>max</th>
<th>sd (( \sigma ))</th>
<th>skewness</th>
<th>kurtosis</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading L1 Thai</td>
<td>138</td>
<td>36.72</td>
<td>26</td>
<td>0</td>
<td>201</td>
<td>34.19</td>
<td>2.35</td>
<td>6.80</td>
<td>0.00</td>
</tr>
<tr>
<td>Reading L2 English</td>
<td>138</td>
<td>10.30</td>
<td>4</td>
<td>0</td>
<td>73</td>
<td>14.73</td>
<td>2.28</td>
<td>5.10</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Table 3.5: Number of pseudowords used to create versions with differing text coverage figures.

<table>
<thead>
<tr>
<th>Target text coverage</th>
<th>Unknown words (pseudowords)</th>
<th>Assumed known words</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>0</td>
<td>633</td>
</tr>
<tr>
<td>98%</td>
<td>12</td>
<td>621</td>
</tr>
<tr>
<td>95%</td>
<td>31</td>
<td>602</td>
</tr>
<tr>
<td>94%</td>
<td>37</td>
<td>596</td>
</tr>
<tr>
<td>90%</td>
<td>63</td>
<td>570</td>
</tr>
<tr>
<td>86%</td>
<td>88</td>
<td>545</td>
</tr>
<tr>
<td>80%</td>
<td>126</td>
<td>507</td>
</tr>
</tbody>
</table>

Proficiency Examination (Hill & May, 1962) which is equivalent to the Proficient User level of the Common European Framework of Reference for Languages. This particular story was chosen for its similarity to graded readers, conventional structure, and simple schema (Hu, 1999).

To further limit the possibility of confounding text variables, word types which are not proper nouns, exclamations, nor found in A General Service List of English Words or Coxhead’s (1998) An Academic Word List were systematically replaced with pseudowords starting with the types with the lowest frequency in the BNC until the desired text coverage figures were achieved. In other words, the word types with the lowest frequency ratings had the highest probability of replacement with pseudowords. An example of the increasing number of pseudowords required to achieve the range of text coverage figures in the original studies and the present study can be found in Table 3.5 (page 71).

Rather than simply reuse the text coverage versions from the original studies, Research Question Two mandates that the recommended 98% text coverage figure be put to the test. Based on the poor results of the 80% text coverage condition in the original studies, it was also avoided in favour of a figure which is some point between that and the next highest text coverage condition (90%). The final text coverage levels were determined by simply taking the distance between 100% coverage and the 98% predicted minimum level, i.e. $100\% - 98\% = 2\%$, and then increasing that distance for each following text coverage level by sequentially increasing the exponent of the first distance. Thus the high text coverage condition was set at 98% $(100 - 2^1)$, the medium was set at 94% $(98 - 2^2)$ and the low text coverage condition was set at 86% $(94 - 2^3)$. Only three conditions were used because the previous studies had already shown that 100% text coverage was ideal and that 80% text coverage was likely too difficult, so the area most in need of further investigation lies somewhere between those two extremes.

---

8Details on the use of pseudowords in this study can be found in subsection 3.3.2, page 72.
Although it may seem tempting to include the 100% text coverage condition for a baseline measure, there is at least one sound reason for avoiding it. First, it appears that the primary source of the “non-homogeneity of variance” (Hu, 1999, p. 78; Hu & Nation, 2000, p. 417) in the reading comprehension measures can be attributed to the lack of variance in the 100% version when compared with the other text coverage conditions. Most statistical tests assume homogeneity of variances, so ignoring the treatment which produces the smallest variance may lead to more balanced groups.

In each of the three versions, inconsistencies discovered in the consistency of pseudowords9 were corrected. Additionally, the length of some pseudowords were changed to more closely match the length of the words which they replaced. The complete source version of the text, along with the three modified treatment versions, can be found in Appendix D (page 193).

3.3.2 Pseudowords

As in the original studies, pseudowords were used rather than low-frequency alternatives such as in Holley (1973). This paradigm is established as a standard method of ensuring that target items are guaranteed to be unknown. For example, pseudowords are often used in studies which investigate the ability to guess unknown vocabulary from contextual clues (e.g. Pulido, 2007; Liu & Nation, 1985). Although incidental vocabulary acquisition is not the focus of this study, the pseudoword paradigm increases the confidence that the text coverage figures of the three treatment conditions are what they are claimed to be.

3.3.3 Basic measures of reading comprehension

The reading comprehension test items were translated into the participants’ L1, Standard Thai and can be found in section E.1 (page 203). This approach to testing L1 reading comprehension in a reader’s L1 is not uncontroversial. There is some empirical evidence that the difficulty of comprehension questions can vary with the language in which they are posed. Filipi (2012), for example, found that written questions for a listening comprehension test in several L2s were generally more difficult when written in the target L2 than in the participants’ L1. Items which measured comprehension of more global or implicit propositions were less difficult in the L1, while questions which had a large amount of overlap with the text, such as a question stem which was verbatim

---

9Details on the inconsistency of pseudowords used to construct the experimental texts can be found in subsubsection 2.7.2.3, page 47.
from the text, were easier to answer in the target L2. Although Filipi doesn’t expound on the possible explanation for this pattern, perhaps because the study was by no means definitive, it seems logical to assume that when an item measures comprehension of an explicit proposition, possibly using some of the same vocabulary, then there would be fewer chances for any information to be lost in translation if the item is in the target L2. In the participants’ L1, however, the participant would need to first recognise that the L1 translation consists of some words from the text before that fact becomes facilitating.

Despite the conflicting evidence, there are strong advocates of testing reading comprehension in a reader’s L1. Bernhardt (2011), for example, is one such researcher who believes that testing should take place in a reader’s L1, but cites no studies to support the belief. Bernhardt’s model of L2 reading includes L1 reading skills as an important factor in L2 reading comprehension, so by extension, a reader’s L1 should be considered when assessing that performance. The other justification is not grounded in theory; Bernhardt advocates using readers’ L1 in assessment because most L2 teachers are themselves non-native speakers of those target L2s so it is more practical to use the language they share with their students.

It probably would have been impossible, and certainly impractical, to construct reading comprehension measures in the variety of L1s of the original study’s participants, but it is fairly simple to do in the present study because of the participants’ shared L1. This avoids the potentially confounding effect of testing in L2 English. Therefore, given that there does not seem to be convincing evidence not to measure reading comprehension in the L1 of the participants, and it seems intuitively obvious that in most cases testing in the L1 would result in a clearer measure of the construct, the participants’ L1 was used in all measures of reading comprehension in this study.

Hu and Nation (2000) used two different protocols to assess reading comprehension: a cued written recall test and a multiple choice question (MCQ) test. Only the MCQ items were used in this study, though not because it is inherently better. The decision was made in deference to Bernhardt who acknowledged that “[t]here is no perfect measure of reading comprehension. Every measure is flawed; each measure provides one perspective. Therefore, multiple measures are necessary to provide a more than unidimensional picture.” (Bernhardt, 1991, p. 224) Since the MCQ and cued written recall tests purport to measure essentially the same construct, and given the limited amount of time for data collection, one of them was dropped from this study to make room for measures of other aspects of reading comprehension to give a more multidimensional
The perspective advocated by Bernhardt. The MCQ test was selected over the cued written recall test for the following three reasons.

The first is simply a practical consideration: evaluating the cued recall responses is somewhat subjective, requires more than one rater, and the increased time and resource investment the evaluation requires yields little, if any, increase in insight. This can be seen in the high correlation between the two tests which was reported to be \( .84, p < .0000 \). (Hu & Nation, 2000, p. 419)\(^\text{11}\) Clearly the high degree of covariance implies that there is not much additional information that can be gained from collecting data from both tests.

Another reason is that the MCQ test data reported in Hu and Nation seem to be more amenable to rigorous analysis. While Hu and Nation acknowledge that the criteria are arbitrary, the MCQ criterion is set firmly at 12 out of a possible 14 whereas the cued recall criterion is set at “around 70” (p. 416) even though the data are aggregated into 10 point bins. There is no explanation for this decision nor why other sensible approaches, such as the midrange of the bin (75) or the upper limit (79) were not chosen.

The final reason is that Hu and Nation seem to focus more on the results of the MCQ test than on the cued recall data which suggests that they felt this measure of reading comprehension made a somewhat stronger case for the conclusions. Although correlations were checked between the subjects’ predictions of performance and their actual performance on the two reading comprehension measures, only the correlations with the MCQ test are reported in detail (see Hu, 1999, Table 3.15 on p. 84 and Table 3.16 on p. 85; Hu & Nation, 2000, Table 6 on p. 420 and Table 7 on p. 421). This is curious because not only were the correlations between the written cued recall scores and the predictions higher, but text coverage also accounted for a greater share of the written cued recall variance (62.18%) than that of the MCQ measure (48.62%). There is more space devoted to reporting the cued written recall measures in the unpublished version compared to the published version, so there must be some reason why the focus in the published study seems to emphasise the MCQ test.

The answer may partially lie in the procedure: the MCQ test was administered before the cued written recall test. Although the effect of this ordering can not be conclusively determined retrospectively, it is possible that the MCQ test served as a sort of rehearsal which is manifest in the increase of explained variance seen in the cued written recall test. P. Johnson (1981) measured reading comprehension with both a written recall and

\(^{10}\)Details of the additional measures of reading comprehension can be found in subsection 3.3.4, page 78

\(^{11}\)The sentence actually reads, “The two comprehension measures correlated with each other at 0.84(\( p > .0000 \))” but it is clear that the inequality symbol was unintentionally reversed.
3.3. MATERIALS

MCQ test in that order. Although P. Johnson’s study is an ANOVA design, the order of test administration was presumably selected to avoid providing additional information about the text in the form of the MCQ prompts. Only a counter-balanced administration of both types of reading comprehension measure could have negated any possible rehearsal effect, though, which neither study did. Nevertheless, the optimal balance between time of administration, validity, and sensitivity point to the MCQ test as the stronger candidate for measuring reading comprehension.

A perennial complaint of MCQ tests is that the limited number of distractors statistically ensures performance of at least chance and most likely higher than that. To decrease these odds, two additional options were added:

- *I remember reading about it but I don’t remember the details*
- *I don’t know*

The first additional option provides the opportunity to register partial knowledge or at least the impression of it. Presumably this sense of knowing that some proposition exists in the text but not being able to fully recount it should occur as text coverage decreases.

The second option captures yet another scenario in reading: not having any memory of the proposition in question at all. It represents the final stage in the degradation of reading comprehension at the propositional level. Just as reading comprehension in general seems to decrease with text coverage, the comprehension of individual propositions might not be a dichotomous, all-or-nothing construct where either complete or incomplete knowledge are the only possibilities.

While these two extra options are not technically distractors which decrease the odds of guessing correctly, they offer alternate responses to participants who genuinely cannot select the correct option with confidence and do not want to adopt a guessing strategy. They provide the opportunity to more fully understand how reading comprehension decreases with text coverage. The original questions can be found in section E.1 (page 203).

The instructions given to the participants prior to reading the text directed their attention to the extra two options and requested that they should select them rather than guess. Not all participants use the same test taking strategies, even when these extra options are available. So to estimate the degree to which participants faithfully followed the test instructions, one unanswerable question, Question Eleven:

Before the policemen came to the house, they were

A) eating dinner.

B) talking with the postman.
C) chasing a murderer.
D) talking with George Hitchcock.

This information required to answer this question can not be deduced or even inferred from the text since the first mention of the policemen in the text is when they appeared at the house.

The idea behind this slightly unorthodox practice is that participants who were resisting temptation to guess in the face of uncertainty, which is a valid test-taking strategy, should respond to the unanswerable question with *I don’t know* or *I remember reading about it but I don’t remember the details*. Failure to do so would indicate a guessing strategy and open up the possibility that chance or some other process may have played a role in the participant’s performance on other items as well. It is possible that performance on this test taking strategy indicator could potentially be useful in analysing the responses as well.

One problem with an approximate replication is the requirement that all materials will be the same, or very close to the same, as the original study. A graph of MCQ test performance in Hu and Nation (2000) shows a clear ceiling effect in the 100% text coverage condition and possibly a slight ceiling effect in the 95% as well. So using *exactly* the same instrument would likely lead to the same problem but changing it too much would prevent this study from being called a replication. In order to remedy this problem, a few supplementary MCQ test items were added to the original pool. The MCQ reading comprehension test in the original study contained nine questions targeting implicit propositions and five questions targeting explicit propositions. The cued written recall test contained a greater proportion of items targeting explicit items, but it was not used in this study. To help balance the ratio slightly in the MCQ test, one additional item, Question One, was created targeting an explicit proposition based on the second question of the cued written recall measure in the original study:

At the beginning, Mrs. Scott was:

A) watching TV.
B) calling her husband.
C) playing with her children.

---

12 Both the unpublished and published versions of the original study state that ten MCQ test items target textually implicit propositions while the remaining four target textually explicit propositions. A count of the explicit and implicit labels next to the questions in the appendix of the published version show that there were actually only nine items targeting textually implicit propositions and five items targeting textually explicit propositions.
D) preparing dinner.

One additional item, Question Ten but labelled as Question Nine in the original study, was reclassified as targeting a textually explicit proposition. The question reads as follows:

The police were called because

A) someone was trying to kill Mrs. Scott.
B) there was a lot of noise.
C) the postman became very worried.
D) Mrs. Scott’s phone did not work.

The part of the text where the answer can be found is

Meanwhile, the telephone-operator, finding I had not put my telephone down and hearing the crashes and knocking, thought there was a fight going on in my house and telephoned the police.

It seems clear that if a reader can equate the words crashes and knocking with the concept of noise, then the proposition is more explicit than implicit.

Although the implicit items still outnumber the explicit items, they are now more closely balanced. This is important because there is some evidence of difficulty differentials between questions targeting implicit propositions compared with explicit propositions (e.g. P. Johnson, 1981). The conditions which cause the differential to arise and their interactions are not well defined. To be on the safe side, then, a balance will ensure there are enough items to analyse explicit and implicit propositional comprehension as distinct constructs.

Two items targeting the text’s genre were added as well to measure the effect of text coverage on scriptually implicit information. After the additions and reclassification, seven items targeted textually explicit propositions, eight items targeted textually implicit propositions, two items targeted scriptually implicit propositions, and one item was technically unanswerable.

Adding items rather than replacing makes it possible to still label this study as an approximate external replication by excluding the supplementary questions in the initial analysis to make a direct comparison with the results in the original study. At the same time, a greater number of items and options increases the chance of greater fidelity when the entire data set is analysed. The addition of the I don’t know and I remember reading about it but I don’t remember the details options could also help reduce the ceiling and floor effects.
3.3.4 Expanded measures of reading comprehension

Alderson (2000) makes the distinction between the \textit{process} and \textit{product} of reading to illustrate one of the obstacles in measuring reading comprehension. While individual readers may make use of various processes when reading, many traditional tests of reading comprehension assume, in practice, that readers will arrive at roughly the same \textit{product} which forms the basis for the questions in a reading comprehension measure. But research shows that readers also arrive at variable products of reading as well. Alderson notes that although there are clearly some reading comprehension products that are clearly wrong, such as those which cannot logically follow from the propositions in the text, who is to decide which product, or products, are \textit{correct}?

It should be fairly uncontroversial to assume that many or most explicit propositions, \textit{i.e.} specific details directly stated in the text, are generally recognised while reading and, ideally, remembered for some time afterwards. Only slightly less controversial should be the notion that inferences within the limited realm of those explicit propositions should also generally be recognised and remembered, perhaps at a lower rate than explicit propositions. Expectations beyond that, however are subject to the idiosyncrasies of readers and the debate of reading researchers.

Any observations of reading comprehension will be theory-laden, as Alderson clearly states:

\begin{quote}
Given the difficulty [...] of the private and silent nature of reading, the individual nature of the reading process, and the often idiosyncratic yet legitimate nature of the products of comprehension, any single technique for assessment will necessarily be limited in the picture it can provide of that private activity. Any one technique will also, and perhaps necessarily, distort the reading process itself. Thus any insight into reading ability or achievement is bound to be constrained by the techniques used for elicitation of behaviour and comprehension. (p. 270)
\end{quote}

One way to cautiously skirt around the issue is to create several open-ended questions which have the potential to give insight into the possible range of reading comprehension products from a given text as well as the variability within each of those products. Open-ended reading comprehension questions were created to represent a range of higher order comprehension domains such as the author’s intended audience, author intent, and even the ability to imagine what might occur after the conclusion of the story. In addition, multi-item scales were created to measure affective responses such as \textit{enjoyment} of the text. Overall, these supplementary questions are exploratory in nature.
and were designed detect a range of possible subtle changes to reading comprehension and related phenomena that changes in text coverage may cause.

The danger in such an approach, as with any open-ended questions, or even the structured cued recall questions used in Hu and Nation (2000), is that evaluation of the responses is highly subjective. Exploratory evaluation of these questions, even if they are not conclusive or significant, may lead to further insights on the relationship between text coverage and reading comprehension above and beyond the unidimensional model presented in the original studies. If something of interest is found, more detailed measurement and analysis can be undertaken in a followup study.

3.3.5 Literacy and reading attitudes

One problem with many reading studies is an unstated assumption that the participants’ core literacy skills are equivalent. Yet Bernhardt’s (2005) model of L2 reading comprehension asserts that a generic literacy ability may account for a significant fraction of reading comprehension variance. Indeed, it is the type of ability which Cummins’s (1976) Threshold Hypothesis and Clarke’s (1980) Short Circuit Hypothesis insist must be blocked from use when a reader’s L2 ability is below the threshold.

In order to add weight to the validity of participant self-reported reading times, the participants’ reading attitudes were measured based on the assumption that more positive attitudes towards reading should correlate highly with actual reading patterns. Just as self-estimates are prone to error, attitude scales are often disproportionately affected by contemporaneous events so it is possible that a usable estimate of both L1 and L2 literacy could be measured by combining the attitude scale and reading times into a composite score.

Measuring literacy in adult learners, however, is not a trivial task. For example, when the National Center for Education Statistics, a part of the United States Department of Education, conducted large-scale measurements of adult literacy under the National Assessment of Adult Literacy (NAAL) programme, several instruments were employed, but the literacy assessment component alone required more than 45 minutes to complete. Additionally, each participant was measured individually (S. White & Dillow, 2005). Even with such a large investment in time, the NAAL does not collect enough information to provide individual scales scores but rather uses the data to es-

\[13\]

Details on the reading times reported by participants can be found in subsection 3.2.5 (page 69).
timate population- and group-level parameters only. To characterise individual ability with great accuracy would require additional items and require even more time.\textsuperscript{14}

In an ideal world, the most precise and accurate instruments would be used for data collection, but compromises must be made. In this case, the principal constraint was time available for data collection, so a shorter multi-item scale would have to be used as a proxy for literacy. The oft-cited McKenna and Kear's (1990) \textit{Elementary Reading Attitudes Survey} (ERAS), consisting of twenty items which measure attitude towards recreational and academic reading in students up to grade six, was considered. Follow-up validation work of the ERAS by McKenna, Kear, and Ellsworth (1995) showed that reading attitudes become more negative as students progressed in their education and the decline is more marked in the least able readers. This suggests that reading attitude may be a useful proxy in explaining performance on a measure of reading comprehension, but the ERAS itself would be unsuitable for use with the young adult participants due to its use of cartoon characters to indicate responses on its Likert scale. Even if the cartoons were replaced with descriptive statements, many of the scale items refer situations which are unique to only very young students such as having reading time in class or reading out loud in class.

According to Summers's (1977) overview of reading attitude instruments, the reading component of Estes's (1971) Estes Attitude Scales (EAS) was, at the time, the best available. Summers praised the EAS for following sound principles in construction and validation, but many of the statements in the scale now seem dated, e.g. “8. Reading is only for grade grubbers” (p. 138), and culturally restricted, e.g. “14. There should be more time for free reading during the school day.” (p. 138) In fact, the problem with most available instruments considered for the present study is that they contain a certain amount of cultural bias which would reduce the reliability and accuracy when translated and used in the participants’ country.

Thai alternatives, however are sorely lacking. The website of the สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน (/sǎm.nák.ŋən kʰá.nаʔ kam.má.kan kam sùk.ʔaː kʰâŋ pǔm.tʰàːn/, Office of the Basic Education Commission) hosts several examples\textsuperscript{15} of questionnaires and surveys purported to measure reading attitudes and habits. None of them, however, make reference to authorship, validation, nor instructions for administering or interpreting results. Most appear to be either direct translations from foreign question-

\textsuperscript{14}The NAAL uses a marginal maximum likelihood IRT model which means that no participant needs to sit the entire test because proficiency is calculated as a probability distribution over all possible scores.

3.3. MATERIALS

Questionnaires or overly simplistic collections of questions without an obvious underlying latent construct.

A compromise was made by selecting a reading attitudes survey designed for roughly the same age group as the participants in this study, even though it was less highly regarded than the venerable EAS, and translating it into Thai. Two existing scales for, one for teens and one for adults, were considered: Tullock-Rhody and Alexander’s (1980) Rhody Secondary Reading Attitude Assessment (RSRAA) and M. C. Smith’s (1990) Adult Survey of Reading Attitude (ASRA). The RSRAA was created to address a perceived lack at the time of reading attitude assessments which were designed to be used with students in years seven to twelve and contained items that accurately reflected the participants’ feelings on reading (Tullock-Rhody & Alexander, 1980). The participants in the present study were first and second year university students, so it was assumed that the scale would be more appropriate to them than the ASRA which was developed and validated on participants in their mid-30s to mid-40s. The RSRAA is also shorter than the ASRA which makes data collection easier.

The RSRAA was developed by collecting statements about people who like to read from a range of American students ($N = 142$) in years seven through twelve representing a range of socioeconomic status. Piloting with similar students ($N = 204$) reduced the number of statements from 33 to 25 which were strongly correlated. Validation was then carried out with another group of students ($N = 349$) where responses on the scale were compared with teachers’ perceptions of students which had the most and least positive attitudes towards reading. A significant difference in mean scores was found between the groups identified by the teachers. Another indication of validity was a high correlation ($r = .84$) between two administrations, one week apart, which suggests temporal stability.

Despite the reasonably rigorous development process, the RSRAA is not an the most ideal candidate for measuring reading attitudes. Although the it has been shown to discriminate between readers with positive and negative attitudes identified by their teachers, the range of scores is not great; the mean differences in extremely positive and negative readers differ by just 15 points although the RSRAA scale has a range of 100 points (25–125). Another limitation of the RSRAA, and attitude scales in general, is the relatively poor correlation between attitude and behaviour. For example, Moore and Lemons (1982) found no significant correlation between stated reading attitudes and selection of reading material. Nevertheless, it is designed for secondary school students, is relatively recent, and doesn’t have any statements which would be odd or strange in a Thai context.
To compensate for the fact that the RSRAA was not created specifically for a Thai context and the questionable ability to predict actual reading behaviour, a quantitative questionnaire which asked participants to estimate the average number of hours spent reading various genres of texts in both Thai and English was also used.\textsuperscript{16} The reported reading times may help provide converging evidence of L1 and L2 literacy, albeit roughly. The reported temporal stability of the RSRAA is an important feature in this regard as the self-reported reading times are assumed to be less accurate. Although this combined approach is probably not as sensitive or as fine-grained as a properly constructed and validated survey specific to Thai college students would be, it may be useful nonetheless.

3.3.6 Vocabulary size

One common problem in generalising across L2 research is a lack of consensus in characterising the participants’ L2 abilities. It is not uncommon to describe participants as members of an intact class all of whom have had five years of compulsory English instruction prior to the study. Even when any kind of ability measurement is mentioned, it is often a broad and unreproducible measure such as performance on a final exam or overall marks in the previous course. Better are studies which report large, standardised test results such as TOEIC, TOEFL, or IELTS scores, but even these are not as useful as they may appear at first blush. Unless the tests were administered as part of the study, it is likely that the scores reported are of varying ages, with no accounting for any learning which has taken place in the interim. Additionally, these tests paint a picture of the learners’ abilities with broad brush strokes which lose the detail needed to characterise participants in a narrowly focused research design.

Measuring participants vocabulary sizes, at the same time as the experiment is conducted, gives the researcher a more principled, reproducible basis for characterising the participants in a study. It is ideal because its measurement is fast, low-cost, and reasonably accurate. Although there have been significant flaws with previous tests of vocabulary size (Thorndike, 1924), standardised tests have begun to appear. The VLT (P. Nation, 1983) has been used by many L2 researchers as a tool to measure vocabulary size, despite the fact that the test is designed as a diagnostic test.

The recently published VST (P. Nation & Beglar, 2007) is designed to measure vocabulary size in both L1 and L2 English (P. Nation, 2012). As Hu and Nation (2000) found no correlation between performance on the VLT and reading comprehension, it may be prudent to replace it with the VST for this replication. If no correlation is expected any-

\textsuperscript{16}For further details and results of the self-reported reading times, see subsection 3.2.5 (page 69).
way, it should not be necessary to administer the VLT to new participants. Replacing the VLT with the VST should therefore not affect the degree to which this study can be called a replication because the substitution would be more similar to an expanded measure than a change.

Capitalising on the common L1 of the participants, the VST, which was recently validated in Beglar (2010), was localised to Standard Thai. For many of the same reasons Bernhardt (2011) advocates testing reading comprehension in a reader’s strongest language, there is reason to believe that vocabulary size should be estimated using a learner’s strongest language to avoid potential difficulty emanating from the questions themselves. The VST manual P. Nation (2012) also recommends using a bilingual version wherever possible. Several localised versions of the VST already exist and there is evidence that using a localised version makes the test more sensitive particularly for lower proficiency learners (Elgort, 2013). Previous localisations have followed a fairly rigorous procedure. Nguyen and Nation (2011), for example, employed two separate L1 Vietnamese users who were also experienced English teachers in the localisation process. One translated the test while the other proofread the translation which was then followed by a joint evaluation of each item. An even more rigorous procedure was adopted by Elgort (2013) in localising the VST for Russian by establishing a clear set of protocols to create consistency in dealing with cognates as well as using two L1 Russian users to proofread the translation.

The Thai localisation of the VST used in the experiment was undertaken before either Nguyen and Nation (2011) or Elgort (2013) were published so the procedure was quite simplistic by comparison. As with other localisations, rather than translating the options literally, the word which each option described was translated as illustrated in Figure 3.1 (page 83). The researcher translated the items into Thai using principles similar to Elgort (2013) with the assistance of one L1 Standard Thai user who was also the participants’ English instructor. However, a thorough proofreading by multiple L1 Standard Thai
users was never performed. Due to data collection time constraints and estimates of the probable range of participant vocabulary sizes obtained from the participants’ English instructor, only the first eight levels, representing 8000 word families, were tested. The localised version can be found in Appendix B (page 183).

3.4 Procedure

The data were collected over a one-week period during the students’ regular class meeting times (Figure 3.2, page 84) in late 2009. On the first day of data collection, participants were informed of the change from the classes’ normal schedule. To ensure that all participants were fully cognisant of the events described in this section, all instructions, both written and oral, were translated to Standard Thai. At the beginning the first day of the procedure for each class, the participants’ regular English instructor introduced the researcher. An information sheet and consent form were distributed to the participants and the researcher gave a general explanation of the purpose of the research which mirrored the content in the information sheet. The participants were given an opportunity to ask questions regarding the research and their participation in it. The researcher emphasised that their performance in the experiment would not in any way affect their marks for the course and that their responses would be anonymised.

After any questions had been answered and consent forms collected, the reading text was distributed to the participants. The researcher read aloud the instructions at the top of the experimental text and confirmed that the participants understood. The MCQ reading comprehension test was exchanged for the experimental text when each participant individually indicated readiness to take the test. The expanded reading comprehension measures were similarly exchanged for the MCQ reading comprehension
test when each participant individually indicated completion. No time limits were imposed for any activity.

The administration of the VST followed a few days later along with a debriefing where the researcher explained the use of pseudowords in some of the texts and provided glossed versions of the original text to prevent the participants from attempting to treat the pseudowords as actual words. The literacy and reading attitudes surveys were completed as homework and therefore, with no time constraint in order to give the participants time to reflect on their true attitudes towards reading in a neutral environment as well as better recollect the time spent on various reading activities during the school break the previous month. The drawback of this approach was that 16 participants neglected to return the survey or misplaced it. Most of the participants did eventually return the completed survey and the data was added to the dataset.

3.5 Analysis

For all analyses in this study, the significance level was set at 0.05 and statistical tests were performed with R (R Core Team, 2014) following the principles of reproducible research. As parametric tests are preferred, all continuous and interval variables were

\[\text{Figure 3.3: Distribution of MCQ scores by observation, by class, and aggregated}\]
evaluated for their shape (kurtosis), breadth (skew), and deviation from normality (Shapiro-Wilk normality test). Dependent variables were checked for homogeneity of variance, an important assumption for regression analysis, using the Fligner—Killeen test.

The analyses were performed twice. The first analysis was carried out to replicate the original studies as closely as possible using the original reading comprehension questions as the dependent variable and text coverage as the independent variable. The results can then be directly compared to the original studies. The second analysis is simply expanded to include the full suite of reading comprehension questions and additional dependent variables to investigate whether other explanatory factors should be added to the model. Exploratory data analysis on the participants’ performance on the MCQ reading comprehension measure, shown in Figure 3.3 (page 85), indicated that the data distribution is approximately normal and there are a range of ability levels between the classes as predicted by their instructor. Classes A and B appear to represent the average ability for all participants, with Class C slightly lower and Class D slightly higher.

3.5.1 Question items from Hu and Nation (2000)

Participants in Hu and Nation (2000) scored at least 14 out of 18, or 77.78% on the 2,000 level and 10 out of 18, or 55.56%, on the 3k level of the VLT (P. Nation & Coady, 1988). In order to match the participants in this study as closely as possible in ability to those in the original study, similar cutoff points can be set for performance on the VST. Although the two tests are not equivalent, by setting a minimum standard of at least eight out of 10 for the 2k frequency level items and at least six out of 10 for the 3k frequency level items, some confidence can be placed in the results. After all, the only way to ensure that the text coverage of each treatment level is actually what it is designed to be is for the participants to know all of the words in the text except the pseudowords. If they don’t also know some fraction of the words in the text, then the predictor variable can not be claimed to be error free which is a fundamental assumption of linear regression.

Tests of conformance to other assumptions of regression analysis were performed. The variances of the original MCQ reading comprehension test items by treatment, as well as a test of heteroscedasticity, can be found in Table 3.6 (page 87). No significant differences were detected so the most critical assumption for regression analysis (Crawley, 2007) appears to be valid.

An ordinary least-squares regression model in the form of $R_c = a + bT_c$ was fitted to the data where $R_c$ is the performance on the MCQ reading comprehension measure and $T_c$ is the text coverage level of the treatment. The fit was evaluated by examining
Table 3.6: Variance in replication MCQ reading comprehension items by treatment group and test of heteroscedasticity under the Strict VST Criterion.

<table>
<thead>
<tr>
<th>Group Variance</th>
<th>MCQ items in Hu and Nation (2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>98% (n = 8)</td>
<td>2.500</td>
</tr>
<tr>
<td>94% (n = 13)</td>
<td>5.900</td>
</tr>
<tr>
<td>86% (n = 9)</td>
<td>1.550</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Test Fligner-Killeen (n = 30)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.075</td>
</tr>
</tbody>
</table>

Figure 3.4: Model fit evaluation plots for Strict VST Criterion
Table 3.7: Variance in replication MCQ reading comprehension items by treatment group and test of heteroscedasticity under the Loose VST Criterion.

<table>
<thead>
<tr>
<th>Group Variance</th>
<th>MCQ items in Hu and Nation (2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>98% (n = 18)</td>
<td>4.260</td>
</tr>
<tr>
<td>94% (n = 21)</td>
<td>3.930</td>
</tr>
<tr>
<td>86% (n = 20)</td>
<td>7.200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Fligner-Killeen (n = 59)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.852</td>
</tr>
</tbody>
</table>

the diagnostic plots found in Figure 3.4 (page 87). The scatter plot of residuals and fitted values showed no discernible pattern indicating the variance is well behaved. Likewise, the Q–Q plot indicates that the data are distributed close to normal and does not raise any concern. No serious threats to the validity of the statistical inferences were apparent. The regression model coefficients can be found in Table 3.8 (page 89) in the column labelled as Strict VST Criterion.

One problem with this initial analysis is that the number of participants who fit the criteria for inclusion is quite low at just 30; the total number of participants is even lower than the 66 in Hu and Nation (2000) despite collecting data from 135 participants, more than a two-fold increase over the original study. By slightly loosening the threshold for inclusion by just one item for both the 2k and 3k frequency levels of the VST, to seven and five out of 10 respectively, the total number of participants can be brought up to 59 which is in much closer agreement with the original study while still maintaining a comparable level of L2 English ability as measured by vocabulary size. Similar to the data used in the Strict VST Criterion, no differences in response variable variance was observed between groups in this larger subset of participants (Table 3.7, page 88) so the same ordinary least squares regression model was fitted to the data. The residuals appeared to be well behaved and the Q–Q plot did not indicate any strong deviance from normal. The regression model coefficients for this subset of the data can be found in Table 3.8 (page 89) in the Loose VST Criterion column labelled.
3.5. ANALYSIS

Figure 3.5: Model fit evaluation plots for Loose VST Criterion

Table 3.8: Linear regression of reading comprehension on text coverage

<table>
<thead>
<tr>
<th></th>
<th>Strict VST Criterion</th>
<th>Loose VST Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-5.4680</td>
<td>-8.0004</td>
</tr>
<tr>
<td></td>
<td>[-19.3663; 8.4303]</td>
<td>[-18.7091; 2.7083]</td>
</tr>
<tr>
<td>coverage.real</td>
<td>16.1885*</td>
<td>18.1389*</td>
</tr>
<tr>
<td></td>
<td>[1.2093; 31.1678]</td>
<td>[6.5794; 29.6984]</td>
</tr>
<tr>
<td>R²</td>
<td>0.1381</td>
<td>0.1423</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.1073</td>
<td>0.1273</td>
</tr>
<tr>
<td>Num. obs.</td>
<td>30</td>
<td>59</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.9495</td>
<td>2.2355</td>
</tr>
</tbody>
</table>

* o outside the confidence interval
3.5.2 Expanded study

The expansion of Hu and Nation (2000) involved supplementing the original predictor variable with others thought to be important to reading comprehension. The variables and a brief description of each are:

- **text coverage**: the ratio of pseudo words to high frequency, assumed known words in the experimental text
- **vocabulary size**: a proxy for L1 English ability as measured by the VST
- **gender**: an unbounded gender category self-identified by the participants
- **Thai reading**: a retrospective self-estimate of the total number of hours per week spent reading a variety of text types in Thai which is a proxy for L1 literacy
- **English reading**: a retrospective self-estimate of the total number of hours per week spent reading a variety of text types in English which is a proxy for L2 literacy

The distribution of vocabulary size did not appear to deviate from normal. However, total hours of reading per week in both L1 Thai and L2 English exhibit marked positive skew. To partially correct for this, the hours were $\log_{10}$ transformed. Descriptive statistics for these three variables can be found in Table 3.9 (page 91). Although gender was technically unbounded, only two genders appeared in the data making it effectively a dichotomous variable. The variances of each of the variables by treatment condition were checked for equivalence and are reported in Table 3.10 (page 92). No problems with equivalence of variances were detected.

A moderate correlation between L1 reading times and L2 reading times was detected at $r = 0.52$. This is to be expected as a manifestation of an underlying, language-independent literacy factor. Vocabulary size is also slightly correlated with reading in both Thai and English, although the correlation is higher with English, which is also not surprising. The correlation is small at just $r = 0.23$. Due to random assignment to treatments, text coverage correlations are essentially zero with all other predictor variables.

A maximal mixed effects regression model was fitted using all predictor variables and all interactions up to four way interactions which allows for the possibility of an interaction between all predictor variables except for one. The participants’ vocabulary size, sex, and reading times in both L1 Thai and L2 English were set as random effects and text coverage was set as the lone fixed effect. Maximal models are highly likely to be filled with redundant and insignificant terms, but they are a good starting point when
Table 3.9: Descriptive statistics for expanded predictor variables.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>mean ($\bar{x}$)</th>
<th>median</th>
<th>min</th>
<th>max</th>
<th>sd ($\sigma$)</th>
<th>skewness</th>
<th>kertosis</th>
<th>Shapiro-Wilk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary Size test</td>
<td>138</td>
<td>35.89</td>
<td>35</td>
<td>16</td>
<td>60</td>
<td>8.83</td>
<td>-0.02</td>
<td>-0.44</td>
<td>0.20</td>
</tr>
<tr>
<td>Weekly Thai Reading Time, ($\log_{10}$ hours)</td>
<td>138</td>
<td>1.42</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0.42</td>
<td>-1.03</td>
<td>2.36</td>
<td>0.00</td>
</tr>
<tr>
<td>English Reading time, ($\log_{10}$ hours)</td>
<td>138</td>
<td>0.76</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.51</td>
<td>0.16</td>
<td>-0.74</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Table 3.10: Variance in expanded predictor variables by treatment group and tests of heteroscedasticity.

<table>
<thead>
<tr>
<th>Group Variance</th>
<th>Expanded MCQ</th>
<th>Vocabulary Size Test</th>
<th>Weekly Reading Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thai</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thai, log_{10}(hours)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>English</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>English, log_{10}(hours)</td>
</tr>
<tr>
<td>98% (n = 44)</td>
<td>8.110</td>
<td>81.420</td>
<td>1602.810</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>323.530</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.330</td>
</tr>
<tr>
<td>94% (n = 47)</td>
<td>9.120</td>
<td>100.710</td>
<td>752.230</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.210</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>97.610</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.180</td>
</tr>
<tr>
<td>86% (n = 44)</td>
<td>9.580</td>
<td>57.070</td>
<td>1189.790</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.140</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>249.980</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.300</td>
</tr>
<tr>
<td>Fligner-Killeen (n = 135)</td>
<td>0.738</td>
<td>0.103</td>
<td>0.003</td>
</tr>
</tbody>
</table>
3.6 Results and discussion

Research Question One asked whether the reading comprehension regression on text coverage coefficients reported in (Hu & Nation, 2000) were the same for data from a different population. Hu and Nation’s (2000) research concluded that regressing L2 English readers’ performance on the MCQ reading comprehension measure on text coverage with the coefficients $\alpha = -17.975$ and $\beta = 0.298$ accounted for $\approx 49\%$ of the variance. In the present study, using essentially the same experimental instruments with a more homogeneous population, both the intercept, $\alpha = -8$, and the slope, $\beta = 0.181$, were both quite different no matter which criterion, loose or strict, is used to determine which data to consider. Although the change in slope ($\beta$) is the most critical figure to consider because it is most central to the predictions made from the model, the other interesting factor to consider is the $r^2$ statistic which determines how much of the variance in the MCQ reading comprehension measure is explained by text coverage. The regression model was found to explain only $\approx 14.23\%$ of the variance which is less than $\frac{1}{3}$ that in the original study.
Although the difference in the explained variance may seem extreme, if taken together, they both comport with Bernhardt’s (2005) compensatory model of second language reading which posits that, for proficient L2 readers, L2 knowledge, of which vocabulary knowledge is a major part, can explain up to approximately 30% of reading comprehension. Indeed, the 30% figure is approximately the average of the variance explained by the Hu and Nation (2000) model and the model used in the present study. Therefore it is not clear whether the variance explained in the original study and the present study are different in the sense that one is correct and the other is incorrect or whether both studies taken together are typical of the range of values one would expect if the true value is somewhere close to Bernhardt’s 30% figure. Given that every measurement is imprecise, it seems reasonable to conclude that the latter explanation is more likely.

The impact of this conclusion is that the significance of Hu and Nation’s (2000) results are inflated. Any model which can explain almost half of the variance in human cognitive performance is significant, but explaining only a quarter or so relegates it to perhaps the level of interesting and significant, but incomplete. It must be noted that text coverage purportedly explained \( 62\% \) of Hu and Nation’s second reading comprehension measure, the written cued recall measure. This figure should be approached with caution because the administration of the two measures was not counterbalanced; all participants completed the MCQ test before the cued written recall test. This is significant for two reasons. First, MCQ test items contain enough information in the items themselves (Katz, Marsh, Johnson, & Pohl, 2001) for participants to routinely score higher than chance on the test even when the text which the questions are based on was never read nor available (Katz, Lautenschlager, Blackburn, & Harris, 1990). The information within the test items can play a larger role in answering than the text (Katz & Lautenschlager, 2001). Although the text was not available to the participants in the original study, it had been read and combined with the information gleaned from the MCQ test items, the participants’ comprehension of the text may have increased. The second reason is that the act of completing the MCQ test, with the items in the same order as the text, could be seen as a sort of rehearsal which also could have increased performance on the cued written recall test.

The Second Research Question asked whether 98% text coverage represents a level at which the majority of readers from a different population can achieve adequate comprehension. Hu and Nation set the criterion for adequate comprehension on the MCQ test as \( \geq 12 \) but the mean for the 98% text coverage treatment under the Strict VST Criterion is just 10.88 and just 9.56 under the Loose VST Criterion. These values are similar to the mean for MCQ test in the 95% text coverage treatment (10.18) in the original study.
Finally, the Third Research Question asked whether the addition of vocabulary size or measures of L1 and L2 literacy as explanatory variables increases the amount of meaningfully explained variance in Reading comprehension. In stark contrast to the original study which found no correlation between performance on the VLT and reading comprehension, the present study found statistically significant relationships between both text coverage and total vocabulary size, as measured by the VST, and reading comprehension. These two predictor variables combined explained \( \approx 36.68\% \) of the reading comprehension variance which is a large improvement over text coverage alone in this particular case and also happens to be within the range which Bernhardt’s (2005) model predicts. Literacy measures were not found to be significant.

As this study was both a replication and expansion of the original study, it should be stated unambiguously that the results do not seem to confirm Hu and Nation (2000). In particular, the model coefficients and the variance the model explains are too different from the original study. That difference could be due to either error in the original study or excessively wide and unreported confidence intervals. The model also appears to be much less powerful at predicting reading comprehension than the data from the original study suggested.

The lack of a clear, unambiguous replication of results, especially compared to correlation studies such as Laufer (1989) and Carver (1994), presents the possibility that the relationship between text coverage and reading comprehension may be spurious. The significant role vocabulary size plays in the expanded study further hints that vocabulary size may be the lurking variable that would explain both text coverage and reading comprehension. It is possible that text coverage is just an artefact of vocabulary size but does not fully encapsulate all of the effects that vocabulary size has on reading comprehension which would make the accuracy of predicting reading comprehension from text coverage highly variable.

### 3.6.1 Limitations

Although this study attempted to improve on the weaknesses in Hu and Nation (2000) while maintaining methodological fidelity, there are still a number of limitations in the present study which must be acknowledged. The decision to strictly follow Hu and Nation’s (2000) methodology by allowing essentially unlimited time to read the text is one of the most significant. In an attempt to increase the ecological validity of the research design, neither Hu and Nation (2000) nor the present study imposed a time constraint on the readers. Indeed, in most language learning settings, and perhaps even most typ-
ical language use scenarios, time is not the primary constraint in reading for pleasure. However, failure to impose a time constraint when researching factors involved in reading comprehensions facilitates the use of a wide range of compensatory strategies which can potentially drown out any signal from the variables under study. Reading behaviour and strategy use are affected by time constraints and such constraints may actually improve reading comprehension (Walczyk, Kelly, Meche, & Braud, 1999) so the lack of one here may have made the relationship between text coverage and reading comprehension harder to detect.

Another limitation is that the attempt to improve the experimental texts by making the pseudoword replacements more consistent may have had a negative impact on reading comprehension. The pseudowords used in Hu and Nation were rather inconsistent and more akin to random noise. The improvement in consistency may have diverted attention away from the task of pure reading towards the related task of guessing unknown words from context. Unknown words encountered only once may be easily ignored, but when encountered throughout a text as different parts of speech or conjugations, it may become alluring to allocate mental resources to ascertaining their meaning.

Finally, reading fluency is another potential confounding variable which should be considered. Most models of reading agree that automaticity or fluency of decoding and lexical access are essential to successful reading. Some degree of fluency may be manifest in the participants’ vocabulary size measurements as the two may go hand in hand, especially at higher vocabulary levels. It would be difficult to attain a large vocabulary if one were not reading, thus the only way to really grow your vocabulary is by reading a lot, and reading a lot will make one a faster reader. If the model were more promising than it appears to be, these factors may be worth investigating further.
Chapter 4

Internal replication of Hu (1999)

Given the unexpected disconfirmatory results of the approximate external replication,¹ an internal replication was conducted to further investigate the reasons behind them. An internal replication is usually performed by a study’s original authors (Porte, 2012) although the possibility of external researchers doing so is a possibility as depositing data in public repositories becomes more commonplace and expected. In the case of Hu and Nation (2000), although the original data sets are no longer available (H.-c. M. Hu, personal communication, February 4, 2010) subject-level summary data are available in the appendices of Hu (1999). It is true that haphazardly applying new models and interpretations significantly increases the chance of committing a TYPE I error, but in this case, it may turn out that some fundamental mistake or oversight was made which could have led to exaggerated effects. Therefore, it would be prudent to take a second look at the data and analyses to ensure there are no egregious errors and that the results were reported accurately. Reevaluation by an independent set of eyes can be informative. If any errors are found, correcting them may lead reveal results which are more parsimonious with the external replication study.

4.1 Research questions

The aims of this internal replication of Hu and Nation (2000) are threefold: to improve the usefulness of the original study by calculating and reporting statistics which are now commonly reported but were not at the time of publication, to verify that the violations of model assumptions were not as significant as claimed, and to verify that all calculations were correct.

¹Details of the external replication of Hu and Nation (2000) can be found in chapter 3, page 59.
The first research question in this study was actually investigated and answered during the preparation phase of the approximate external replication, but it is reported here because it is more thematically appropriate:

**Research Question One:** What are the effect sizes and confidence intervals of the linear regression models reported in the original studies?

Reporting these statistics is becoming standard in most scientific fields and are particularly useful when reporting on relationships between variables. They make it possible for the research to be compared with replications and similar studies, so it is vital that they are calculated and reported. When preparing for the approximate external replication, these values informed the researcher of the possible range of values to expect.

The other important figures from the original study are the model coefficients themselves. Hu and Nation (2000) warn that their chosen model, a “simple regression” (Hu & Nation, 2000, p. 417),\(^2\) violates at least three assumptions. These violations are:

1. the variables are discrete not continuous
2. coverage of known words (the predictor variable) has an allowed maximum, and
3. there is non-homogeneity of variance (Hu & Nation, 2000, p. 417).

Statistical tests are robust to some violations yet extremely sensitive to others. Hu and Nation dismiss the violations as unimportant because “both analyses broadly speaking are consistent with each other” (p. 417) but this is not sufficient reason. The lack of homogeneity of variance, for example, is troubling since it is the “most important assumption” (Crawley, 2007, p. 291) for a regression analysis. Therefore, the second research question in this study is:

**Research Question Two:** Did any of the linear regression assumptions which were violated in the original analyses have any effect on the model coefficients?

Finally, when the wide range of limitations identified in the original study is juxtaposed against the high level of credibility given to this study, as evidenced by the sustained high citation rates, it seems reasonable to recalculate all of the statistics presented in Hu and Nation (2000). Doing so would verify that there are no other lurking errors. The third research question will therefore be:

**Research Question Three:** Were there errors in any of the calculations performed in the original studies?

\(^2\)It is assumed that the “simple regression” (p. 417) reported in Hu and Nation (2000) refers to an ordinary least squares linear regression.
4.2 Original data evaluation

Both the original unpublished Master’s thesis (Hu, 1999) and the published study (Hu & Nation, 2000) feature tabular summaries of the reading comprehension measures by text coverage treatment. For the MCQ measures, the original data can be reconstructed from the table but the table of the cued recall measures contains only aggregated data. Fortunately, Appendices 23–26 in Hu (1999, pp. 163–166) consist of a subset of participant data in tabular form. These data are:

- total unweighted MCQ reading comprehension score,
- total cued written recall score,
- estimate of per cent of story understood,
- prediction of performance on a 15-item MCQ reading comprehension test,
- total number of correct items on the vocabulary levels test,
- and total number of words in the cued written recall measure.

In order to verify the accuracy of these data, the summary statistics for each measured variable were calculated for each text coverage condition and compared with the figures reported in both studies. The results of this comparison can be found in Table 4.1 (page 100) which shows that some of the descriptive statistics reported in the original studies can not be reproduced from the data in the appendices of Hu (1999). For example, the mean in the 95% text coverage condition for the cued written recall scores in Table 5 (Hu & Nation, 2000, p. 416) is 61, but when calculated from the appendices data, the mean is 60 when participants with missing data are excluded (listwise deletion). This should be the case based on the description of the statistical analyses, but when all participants are included and missing values are handled by pairwise deletion, (only the missing data points are excluded), the recalculated mean then matches that reported in the table.

A closer examination of Appendix 24 in Hu (1999, p. 164) shows that three data are missing for participants B17 and B18 but, unlike two of the missing data, one is not explicitly marked as missing. Comparing the contingency tables in the original studies with the ones recreated here (Table 4.1, page 100) it appears that there are some minor discrepancies. There are four cells in Table 4 (Hu & Nation, 2000, p. 415) and two cells in Table 6 (p. 420) which do not match contingency tables created from data in the appendices. The conclusions that can be drawn from these discrepancies is that it is impossible to precisely reconstruct the original data from the unpublished and published studies due to the existence of multiple minor errors in reporting the data. The discrepancies are not large, certainly not large enough to preclude an internal replication, but
Table 4.1: Comparison of original and replication summary statistics.

<table>
<thead>
<tr>
<th>Text coverage</th>
<th>n</th>
<th>( \bar{x} ) (mean)</th>
<th>( \sigma ) (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQ Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>17</td>
<td>17</td>
<td>12.24</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
<td>17</td>
<td>10.18</td>
</tr>
<tr>
<td>95</td>
<td>16</td>
<td>16</td>
<td>9.50</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>16</td>
<td>6.06</td>
</tr>
<tr>
<td>Cued Recall Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>17</td>
<td>17</td>
<td>77.17</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
<td>17</td>
<td>61.00</td>
</tr>
<tr>
<td>95</td>
<td>16</td>
<td>16</td>
<td>51.31</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>16</td>
<td>24.62</td>
</tr>
<tr>
<td>Cued Recall Words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>17</td>
<td>17</td>
<td>245.64</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
<td>17</td>
<td>200.11</td>
</tr>
<tr>
<td>95</td>
<td>16</td>
<td>16</td>
<td>175.50</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>16</td>
<td>142.50</td>
</tr>
<tr>
<td>Prediction Correct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>17</td>
<td>17</td>
<td>11.41</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
<td>17</td>
<td>9.65</td>
</tr>
<tr>
<td>95</td>
<td>16</td>
<td>16</td>
<td>7.81</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>16</td>
<td>6.00</td>
</tr>
<tr>
<td>Per Cent Understood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>17</td>
<td>17</td>
<td>90.88</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
<td>17</td>
<td>82.06</td>
</tr>
<tr>
<td>95</td>
<td>16</td>
<td>16</td>
<td>64.38</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>16</td>
<td>47.81</td>
</tr>
<tr>
<td>Vocabulary Score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>17</td>
<td>17</td>
<td>63.82</td>
</tr>
<tr>
<td>90</td>
<td>17</td>
<td>17</td>
<td>64.65</td>
</tr>
<tr>
<td>95</td>
<td>16</td>
<td>16</td>
<td>59.06</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>16</td>
<td>66.75</td>
</tr>
</tbody>
</table>
4.3 **CONFIDENCE INTERVALS AND EFFECT SIZES**

<table>
<thead>
<tr>
<th></th>
<th>MCQ</th>
<th>Cued recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>$-17.9298^*$</td>
<td>$-181.3351^*$</td>
</tr>
<tr>
<td></td>
<td>$[-24.8521; -11.0075]$</td>
<td>$[-226.4274; -136.2427]$</td>
</tr>
<tr>
<td>coverage</td>
<td>0.3005$^*$</td>
<td>2.5737$^*$</td>
</tr>
<tr>
<td></td>
<td>$[0.2250; 0.3759]$</td>
<td>$[2.0821; 3.0652]$</td>
</tr>
</tbody>
</table>

$^*$ o outside the confidence interval

Table 4.2: Reading comprehension regression on text coverage coefficients from Hu and Nation (2000) with added confidence intervals.

their existence does add some uncertainty to any conclusions that can be drawn. As a general rule of thumb, it seems prudent to place the greatest confidence in the descriptive statistics simply because there are fewer of them and hence a lower chance of containing a data entry error. Where participant-level data is required, such as when answering Research Question Two, there is no alternative but to rely on the data in the appendix of Hu (1999) with the hope that any errors are inconsequential.

### 4.3 Confidence intervals and effect sizes

Every measurement contains error. Acknowledging that fact, most statistical results are now presented with confidence intervals which indicate the range in which the true value of a parameter may be found. Unfortunately, Hu and Nation (2000) do not report confidence intervals but this is not surprising because journals which publish research in applied linguistics have only recently begun to require it. Without confidence intervals, however, we have no real inkling of the accuracy of regression model coefficients, *i.e.* the intercept ($\alpha$) and the slope ($\beta$), reported in the original studies. The slope parameter is perhaps the most important piece of information in Hu and Nation (2000) because it is the figure that describes the relationship between text coverage and reading comprehension. The change in reading comprehension that can be predicted by a change in text coverage is captured in the slope coefficient.

To answer Research Question One, the confidence intervals are reported in square brackets below their respective coefficients in Table 4.2 (page 101). A graphical representation, with confidence intervals added as the shaded area about the regression line, can be found in Figure 4.1 (page 102) for the MCQ reading comprehension measure.
Figure 4.1: Replication of plot from Hu and Nation (2000) with added confidence intervals for MCQ reading comprehension measure regressed on text coverage.

Figure 4.2: Replication of plot from Hu and Nation (2000) with added confidence intervals for cued written recall reading comprehension measure regressed on text coverage.
and in Figure 4.2 (page 102) for the cued written recall reading comprehension measure. Adding the confidence intervals to the original regression model coefficients and plots of the original data and regression lines gives some perspective into the degree of uncertainty involved. The wide range of possible values casts serious doubt on the conclusion that such a precise figure as 98% text coverage would result in “adequate unassisted comprehension of the text.” (Hu & Nation, 2000, p. 422)

One other important piece of information which can be calculated from the reported figures should be mentioned directly. The $r$ statistic, more formally known as Pearson’s $r$ or Pearson’s product-moment correlation coefficient, communicates the direction of the relationship as well as the magnitude and is essential for meta-analyses and sample size planning. The $r$ statistic is related to the $r^2$ statistic, the latter obviously being the square of the former. The $r^2$ statistic expresses the proportion of variance explained by a regression. Therefore, to complete the statistical reporting in Hu and Nation (2000), and finish answering Research Question One, the effect sizes can readily be calculated by taking the square root of the $r^2$ statistics originally reported: for the MCQ reading comprehension regression, the effect size is $\sqrt{0.4876} = 0.7$ and $\sqrt{0.6220} = 0.79$ for the cued written recall regression. Both effect sizes are relatively large which indicates that text coverage is likely to be a significant factor in predicting reading comprehension.

### 4.3.1 Testing assumptions

Among the three assumption violations listed in the original studies, the violation of the assumption of homogeneity of variances, known as heteroscedasticity, is the most serious. Hu and Nation do not specify which variables exhibit heteroscedasticity nor to what degree, so it is impossible to know which conclusions might be affected by this violation. Using the participant data from the appendices in Hu (1999), the variance of each measurement for each text coverage treatment group is presented in Table 4.3 (page 104) along with the Fligner-Killeen test of heteroscedasticity which tests the null hypothesis that there are no differences in variance between groups. A significant result was found for both the MCQ reading comprehension measure and the cued written recall quantity of words measure. The latter is not reported in the published study so it is likely that the results were not as compelling to Hu and Nation. The MCQ reading comprehension measure, however, holds pride of place as being the first part of the results section. Further investigation of the significance of this violation on the reported results is therefore warranted.
Table 4.3: Variance in Hu (1999) independent variables and tests of heteroscedasticity.

<table>
<thead>
<tr>
<th></th>
<th>MCQ Score</th>
<th>Cued Recall Score</th>
<th>Cued Recall Words</th>
<th>Prediction Correct</th>
<th>Per Cent Understood</th>
<th>Vocabulary Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group Variance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100% ($n = 16$)</td>
<td>0.970</td>
<td>209.90</td>
<td>2262.650</td>
<td>3.880</td>
<td>85.110</td>
<td>162.780</td>
</tr>
<tr>
<td>95% ($n = 16$)</td>
<td>9.280</td>
<td>329.88</td>
<td>3204.110</td>
<td>5.390</td>
<td>285.950</td>
<td>84.890</td>
</tr>
<tr>
<td>90% ($n = 18$)</td>
<td>7.600</td>
<td>254.90</td>
<td>1896.670</td>
<td>4.960</td>
<td>486.250</td>
<td>141.000</td>
</tr>
<tr>
<td>80% ($n = 17$)</td>
<td>3.800</td>
<td>124.38</td>
<td>1913.050</td>
<td>7.070</td>
<td>789.900</td>
<td>168.330</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fligner-Killeen ($n = 67$)</td>
<td>0.005</td>
<td>0.18</td>
<td>0.537</td>
<td>0.838</td>
<td>0.017</td>
<td>0.514</td>
</tr>
</tbody>
</table>
Another assumption violation is that the variables are discrete, not continuous. This claim seems to be only partly true; it only applies to the dependent variables, not text coverage, the independent variable. Text coverage is continuous between the limits of zero and 100, but for the purposes of the study, only specific points on that continuum were measured. Although the dependent variables are discrete, many of them consist of a large number of possible values such that they start to behave like continuous variables. The cued written recall score and total words are examples of this phenomenon. Likewise, the MCQ reading comprehension measure comprises only 14 questions, but there is no reason why it could not be longer especially if the text were longer as well. The number of questions and the length of the test are artefacts of the need to complete the data collection within a reasonable amount of time, so selecting a different type of regression more appropriate to discrete values would likely improve the fidelity of the model only marginally at the expense of reducing its generalisability to other reading comprehension measurement situations. This violation can therefore be seen as minor unless more compelling evidence to the contrary is found.

Finally, the allowed minimum and maximum values are indeed problematic. Looking at plots of the data, Figure 4.1 (page 102) for the MCQ reading comprehension measure and in Figure 4.2 (page 102) for the cued written recall reading comprehension measure, it seems that part of the source of heteroscedasticity might be caused by a ceiling effect. There also appears to be a slight floor effect as well: the tests not only have allowed minimums of zero, but the MCQ test also a practical minimum of around four as that would be the expected score of a participant guessing at random.

Both variables are strictly bounded; there is no conceivable scenario where one could encounter text coverage greater than 100% or less than zero. Likewise, there are a fixed number of questions in an MCQ reading comprehension measure, so there is no way to achieve a level higher than the maximum number of questions or, again, below zero. But this is exactly the situation the Hu and Nation’s (2000) simple linear regression leads to with the negative intercept of $-17.975$ for the MCQ and $-181.335$ for the cued written recall reading comprehension measures.

In an ideal world, these constraints could be considered as proportional data and handled effectively with a generalised linear model to perform a logistic regression with binomial errors (see Crawley, 2007, chapter 16). Due to the limited range of text coverage values, such a sophisticated model would likely not represent the data well anyway. So, as with the problem of discrete and continuous variables, the nature of the data imply that a more sophisticated model would not be profitable despite the assumption violations.

It appears that only the heteroscedasticity in the data might truly be troublesome. To completely answer Research Question Two, the same ordinary least squares linear regression was performed on a subset of the data in the Hu (1999) appendices. After
Table 4.4: Comparison of original and revised MCQ regressions

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>MCQ_score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
</tr>
<tr>
<td>coverage</td>
<td>0.300**</td>
</tr>
<tr>
<td>(0.039)</td>
<td>(0.060)</td>
</tr>
<tr>
<td>Constant</td>
<td>-17.930*</td>
</tr>
<tr>
<td>(3.532)</td>
<td>(5.324)</td>
</tr>
</tbody>
</table>

| Observations            | 66        | 49          |
| R²                      | 0.488     | 0.320       |
| Adjusted R²             | 0.480     | 0.306       |
| Residual Std. Error     | 2.306 (df = 64) | 2.624 (df = 47) |
| F Statistic             | 60.911*** (df = 1; 64) | 22.126*** (df = 1; 47) |

Note: *p < 0.1; **p < 0.05; ***p < 0.01

Figure 4.3: Revised Hu (1999) model fit evaluation plots.
removing the 100% text coverage treatment data, the Fligner-Killeen test failed to reject the null hypothesis \((p = 0.26)\) that the variances are homogeneous. Therefore, the assumption of homogeneity of variances is no longer violated. A new ordinary least squares regression was performed and the coefficients compared with those in the original study. The results can be found in Table 4.4 (page 106). The fit of the model can be evaluated from the plots found in Figure 4.3 (page 106) which shows a slight departure from normality in the Q–Q plot in the lower left and a subtle arc in the plot of the data. A graphical representation of the regression model along with its confidence intervals can be found in Figure 4.4 (page 107).

Perhaps the most striking observation to be made in comparing the revised model with original model is not the fairly small difference in model coefficients, but rather in the substantial decrease in the amount of MCQ reading comprehension variance explained by text coverage. The effect size remains large, but much less convincingly so. Interestingly, the amount of variance explained by the revised model is perfectly in line with Bernhardt’s (2005) model of L2 reading which allocates about 30% of reading comprehension explainable variance to L2 proficiency. The answer to Research Question Two, then, is that the model coefficients did not seem to be drastically affected by the assumption violations, but the effect size was dramatically reduced. This suggests that
the violation was too big to simply ignore in the original analysis and should have been addressed directly.

### 4.4 Error checking

Hu and Nation (2000) actually report two models of reading comprehension: one for MCQ tests and the other for cued recall tests. The coefficients for both of the models were found by “fit[ting] a model to the data where on average an increase in coverage of vocabulary led to a predictable increase in comprehension.” (p. 419) As the replication study used only the MCQ test, that model will be the focus of this validation.

The regression coefficients for the MCQ test are a slope coefficient of $\beta = 0.298$ and an intercept of $\alpha = -17.975$. A linear regression takes the form

$$ Y = \alpha + \beta x. $$

Substituting the coefficients reported by Hu and Nation (2000), the model can be written as

$$ R_c = -17.975 + 0.298T_c $$

where $R_c$ is expected score on a reading comprehension test consisting of fourteen multiple choice questions similar to those in Hu and Nation (2000) and $T_c$ is the text coverage of a narrative fiction text from which the questions are drawn.

Using this formula, Hu and Nation (2000) conclude that “98% coverage would have yielded an average score of 11.53 on the multiple choice test” (p. 419). Since the scores are discrete, e.g. less than whole number marks are not allowed, 11.53 is rounded up to 12 which is the minimum “arbitrary level” (p. 414) set by the researchers as indicative of adequate reading comprehension.

There is a flaw, however, in this conclusion which becomes apparent when the model coefficients are used to verify the results. Substituting 98% (0.98) for $T_c$ in the model above, we find that

$$ 0.298 \times 0.98 - 17.975 = 11.229, $$

which must be rounded down to 11, not up to 12, due to the discrete nature of marks on a multiple choice test. It is curious that such a simple calculation could result in an error, so investigating from whence the 11.53 figure came may shed light on the source and severity of the error.
Table 4.5: Difference between calculating reading comprehension scores as real and discrete values.

<table>
<thead>
<tr>
<th>Text Coverage</th>
<th>Predicted reading comprehension score (real values)</th>
<th>Predicted reading comprehension score (discrete values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>11.83</td>
<td>12</td>
</tr>
<tr>
<td>99%</td>
<td>11.53</td>
<td>12</td>
</tr>
<tr>
<td>98%</td>
<td>11.23</td>
<td>11</td>
</tr>
<tr>
<td>97%</td>
<td>10.93</td>
<td>11</td>
</tr>
<tr>
<td>96%</td>
<td>10.63</td>
<td>11</td>
</tr>
<tr>
<td>95%</td>
<td>10.34</td>
<td>10</td>
</tr>
</tbody>
</table>

Substituting 11.53 for $R_c$ in the model,

$$11.53 = 0.298T_c - 17.975,$$

and solving for $T_c$,

$$\frac{11.53 + 17.975}{0.298} = 99.01,$$

it appears that Hu and Nation (2000) have somehow misreported text coverage level which would give most readers a chance at achieving adequate reading comprehension. It is difficult to infer how or why the mistake happened, but it is clear that an incorrect result has been reported. According to the model, then, the conclusion should be that a text coverage of at least 99%, not 98%, is required to achieve adequate reading comprehension. Table 4.5 (page 109) demonstrates the error and the corrected conclusion clearly by showing the discrete reading comprehension scores predicted by the model between text coverage levels of 100% and 95%. It shows that text coverage levels between 99% to 100% will likely yield a score indicative of adequate reading comprehension, but that 98% is really no different to 96% text coverage. The range of 96% to 98%, based on the model coefficients, will yield a score of only 11, which was determined to be inadequate.

Research Question Three asked whether there were errors in any of the calculations performed in the original studies. The errors in calculating the ideal text coverage needed for adequate reading comprehension when reading for pleasure demonstrate that the conclusions in Hu and Nation (2000) are incorrect and should be revised. The error also invalidates a number of studies which are based on that figure such as I. S. P. Nation (2006), Webb and Macalister (2013), and Kaneko (2013).
CHAPTER 4. INTERNAL REPLICATION OF HU (1999)

4.5 Conclusion

This brief study reexamined the data from Hu (1999) and Hu and Nation (2000) in an attempt to fill in gaps in statistical reporting which were not standard at the time of original publication. It also investigated violations of the assumptions for the ordinary least squares regression analysis in greater depth than originally reported to determine the effect they might have on the veracity of the reported results. Finally, as far as practical, the original statistical tests were repeated to ensure no other mistakes had been made.

It may appear unwarranted that such a level of scrutiny would be applied to a published study which has already undergone peer review. While internal replications are exceedingly rare in applied linguistics, and perhaps even rarer when carried out by a third party, the original justification for performing an approximate external replication and expansion, combined with the attenuated signal found in those results, applies with equal felicity to this internal replication. The importance of the study justifies a meticulous examination. If no mistakes had been uncovered, as one might wish of every published study, then researchers could build upon it further with confidence.

Unfortunately, serious errors were found which cast doubts on the effect size of the relationship and the accuracy of the model. As described in section 2.5 (page 31), Hu and Nation (2000) is essentially the only published study to establish a causal relationship between text coverage and reading comprehension. Although this internal replication and the approximate external replication and expansion in chapter 3 (page 59) both support the hypothesis that there is a causal relationship, the size of the effect is much less compelling than originally thought.
Chapter 5

Critical evaluation of the text coverage model

We give our highest rewards to those who convincingly disprove established beliefs.

Sagan, *The demon-haunted world*, p. 33

In the conclusion to an article summarising text coverage and its uses, Laufer (2013) advises that “a text’s lexical profile and learners’ vocabulary size, for which quantitative measures are available, provide teachers with the necessary information about the lexical ease and difficulty of different texts for different learners.” (p. 871) The statement gives the impression that the relationship between text coverage, as calculated by comparing a text’s lexical profile and a learner’s vocabulary size, and reading comprehension, couched as *lexical ease and difficulty* by Laufer, is a robust phenomenon for which there is sufficient understanding and replication to justify summarising it in a regular column of a well-respected journal associated with a large professional organisation of L2 English teachers.

It is unclear, however, what the information provided by these measures actually is. Earlier in the same article, Laufer cites evidence that L2 reading requires a certain threshold of L2 ability and that L2 vocabulary knowledge in particular could be the best predictor of reading proficiency. The implication is that L2 vocabulary, or lack thereof, is the primary determinant of whether the threshold of L2 ability has been met and is thus the gateway to L2 reading comprehension. If this is the case, then text coverage, or a “lexical threshold” (Laufer, 2013, p. 868) represents a model of reading comprehension.

A *model* is a simplification of a system or process, such as reading, for some purpose, such as predicting reading comprehension. The intended purpose or use of the model is
what defines its scope and level of complexity. As this model seems to be used primarily for prediction, its scope and complexity are limited, but those characteristics should not have a negative impact on the model’s skillfulness. That is, the model must be able to make accurate predictions. Laufer (2013) asserts that the 98% figure attributed to Hu and Nation (2000) is the optimal threshold for comprehending “authentic written texts” (p. 869). The replication study in chapter 3 (page 59) and the errors uncovered in chapter 4 (page 97), however, cast serious doubt on the veracity of that claim.

As promised in the Preface to this thesis, the structure of this thesis reflects the circuitous journey researchers sometimes must make in their quest for the truth. This chapter represents that journey; it attempts to go back to first principles and axioms with fresh eyes and a sceptical mind. Rather than cast aside the internal and external replication results in this thesis, perhaps attributing the results to some methodological error and trying again, it may be useful to critically reflect on the evolution of the text coverage model itself and the evidence which supports it to determine whether there may be even deeper reasons to question the model. This chapter will focus specifically on the following areas:

- possible explanations for the unexpected replication results,
- the manner in which studies supporting the text coverage model have been understood and summarised,
- the ways important terms have been operationalisation and the methodologies used to investigate them,
- and an investigation as to how the model has evolved.

5.1 Possible explanations

The results of replication studies, especially in the social sciences, are expected to vary. The results of the external replication and expansion compared with Hu and Nation’s (2000) results, after correcting for the errors uncovered in the internal replication, are not as dramatically different as originally thought. They are surprising, though, and the effect size is still much smaller than the revised model. What might explain the decline in effect size between the two?

One explanation is regression to the mean. This is one of the reasons why replication studies are so important and the main justification for conducting a meta-analysis. A single measurement will not represent the true value. A commonly cited and well known example of this phenomenon are athletes who perform exceptionally well in their sport in a particular year. It is almost impossible to maintain that level of performance and
they are virtually guaranteed to exhibit lacklustre performance in the following year. The athletes’ extremes in performance are noted for their rareness, but the fact that they are outliers means that their continuance is untenable. Given time, their average performance will return to more normal levels, *i.e.* regress to the mean.

In a research context, unique or surprising results are seen as fodder for new theories even if they are as much of a fluke as an athlete’s brief streak of impressive performance. The probability of measuring the true value of a parameter increases with the sample size. Consequentially, studies which generate relatively small datasets are particularly vulnerable to spurious results because as sample size increases, the margin of error decreases (Ioannidis, 2005b; Ioannidis, 2005b).

Large-scale studies, especially those involving humans are time-consuming and expensive, so it seems acceptable to begin by collecting data from a small number at the outset. When interesting results are detected, additional data could be collected to verify the results. If regression to the mean is to be expected, then it should be no surprise that subsequent datasets show a more modest effect. This is likely what has happened here.

Text coverage does appear to have a measurable effect on reading comprehension, but the true effect size does not appear to be as large as that reported in Hu and Nation (2000). Another possibility is publication bias. Kurnia (2003), a doctoral study conducted just a few years after Hu (1999) at the same institution and under the same supervision, remains unpublished despite employing a strikingly similar methodology yet concluding that unknown word density accounted for just 2% of reading comprehension variance, even lower than the replication and expansion study presented in this thesis. This phenomenon is known as the *File Drawer Problem* (Rosenthal, 1979) which asserts that, in the most extreme case, about five percent of journal articles are actually false due to *type I* errors. The research which remains in researchers’ file drawers, unpublished, is the 95% of research which failed to reject the null hypothesis. The cumulative effect of this problem is the publication of studies purporting to show an effect even though the same effect may have been studied many times over already. The unpublished studies all failed to reject the null hypothesis, presumably because that is the correct state of affairs and no real effect exists. If a study is repeated often enough, chance will ensure occasional contrary results.

Fukkink, Hulstijn, and Simis (2005) is an interesting example of how possibly false research was almost published. In a study of the effect of lexical access training on higher-order reading comprehension, Fukkink et al. performed two experiments, the latter of which is of interest here. The research design was a counterbalanced $2 \times 2$
Latin square. Each group was trained on half of the words from a common pool, but the tests at the conclusion of the training included all of the words from the pool. This allowed each group to serve as a control for the other, and in a broader sense, it could be seen as both an experiment and a concurrent replication.

The post-treatment measures included, among others, reaction times on a lexical decision task for two groups of words: familiar, which were likely to already be known, and unfamiliar, which were not likely to be known and thus the independent variable. Theoretically, the results should have been roughly the same, but they weren’t. Had the design only included Group A, one conclusion of the study might have been that the training improved the reaction times of the familiar words but not the unfamiliar words. This is a slightly counter-intuitive result which may have led the researchers to conclude that something had gone wrong and perhaps kept the results in Rosenthal’s (1979) metaphorical file drawer. Group B’s results, however, showed the exact opposite effect; the reaction times of the unfamiliar words increased after training. The effects of the training on other results were either small or ambiguous leading the authors to conclude, despite partial support for their hypotheses, that the training had little effect other than increasing word recognition times, albeit inconsistently. It is a fact that the authors did not dwell on.

The uncomfortable truth is that, for various reasons, the field of applied linguistics lacks initiative in conducting replication research and is even likely slightly biased against it. Researchers and journals are more interested in creating new models and theories, which is understandable in a relatively new field of research, than in accumulating evidence about existing ideas. Even in more mature domains such as medicine, as Freedman (2010) so eloquently writes, “it is far more rewarding to add a new insight or exciting-sounding twist to existing research than to retest its basic premises” (para. 11). But researchers and practitioners must keep in mind, both when conducting research and reading published research, that the publication of one study, or even a few studies, is not enough to establish a theory or characterise a phenomenon. The process of arriving at a conclusion requires much more than that, as clearly stated by Novella (2010): “[c]onsensus and reliable conclusions are built upon multiple independent lines of evidence, replicated over time, all converging on one conclusion.” (“The Decline Effect”, para. 3)


5.2 Summaries of previous research

In order to advance knowledge, we must stand on the shoulders of giants; researchers can not make progress without relying on the results of those who came before them. Current knowledge is built up in layers where the lower layers are theoretically more reliable because they have been studied more extensively and are more well documented. The uppermost layers, representing the newest research results, are the most tenuous not only because they are new, but also because the intermediate layers upon which they are built are still subject to revision and even falsification. With replication and further research, these layers become more solid and reliable as foundations for further research while those which don’t stand up to scrutiny are eventually rejected.

This dynamic process is extremely vulnerable in fields such as in applied linguistics where the formal expression of ideas is not common. Much like the problems of operationalising key terms without precision, the same holds true for the summarising and reporting of previous research. Without a tradition of formal expression of terms and the relations between them, previous research in text coverage studies tends to be the victim of careless recasts and synonym substitutions which, at best, obfuscates and, at worst, misrepresents the original findings. An in-depth examination of three highly cited studies representing a range of depths in the layer analogy will make this point clear.

5.2.1 M. West (1926)

As noted in section 2.2 (page 18), many text coverage studies reference the work of M. West. Holley (1973), for example, notes that “Michael West (1926, 1929) held that in teaching English to Bengalis the optimum [rates of vocabulary introduction for a given instructional period] was one new word per fifty running words and that one new word per thirty known words would be preferable for French to English.” (pp. 340–341) There is no entry for M. West (1929) in the list of references, but presumably the citation is referring to M. West (1927) which is. This seems to be a fairly accurate recast of M. West’s recommendations, especially in the light of Holley’s attempt to empirically test those claims.

The reporting of the results began to shift from a recommendation for introducing new vocabulary to a predictor of reading comprehension. Chujo’s (2004) understanding of M. West’s recommendation is a clear example: “Historically, experienced teachers such as West (1926) considered one unknown word in every fifty words to be the minimum threshold necessary for the adequate comprehension of a text.” (p. 231) Although
the text coverage recommendation was a compromise intended to maximise opportunities for vocabulary acquisition while avoiding problems in comprehension, M. West did not relate it with adequate comprehension.

Precision in summarising previous research is critical not only when describing the theoretical framework for one’s own research, but for future researchers as well because the description of previous research can colour future interpretations. Indeed, sometimes so much trust is placed in the summaries that primary sources become irrelevant as can be seen in Hsu’s (2009) dependence on Chujo’s citation alone: “West (1926) considered ‘one unknown word in every fifty words’ to be the minimum threshold necessary for the adequate comprehension of a text (cited in Chujo, 2004, p. 231).” (p. 44)

One aspect of M. West’s text coverage recommendations which seems to get lost in most citations is the context in which they were made. M. West was a staunch advocate of restricting text coverage but not for the purposes of reading comprehension. The purpose was to optimise vocabulary acquisition. In Bilingualism (M. West, 1926/2003), the source of the material in M. West (1926), the following criteria are considered when constructing texts for use in M. West’s experiments on reading:

A. The number of new letters introduced in the First, Second, Third, Fourth hundred words.
B. The number of words in the vocabulary when the first narrative is introduced.
C. Running words per new word.
D. Avoidable words of low frequency as a percentage of vocabulary.
E. Synonyms as a percentage of the vocabulary.
F. Mean Credit-index
G. Zero words (net)
H. Number of new words
I. Length (running words) (p. 275)

In finding a balance among these criteria, the resulting texts were carefully crafted and could hardly be labelled as authentic. Carr (1934) gives a clear description of the types of text that M. West had constructed by describing 12 French texts which had been created in accordance to M. West’s principles of materials development:

Each story begins with the statement “You now know .......... words.” New words are then introduced at the rate of only three or four to a page. These words are immediately repeated several times in the text and also become the objects of special drill at the end of the story in which they are thus
functionally introduced. The meaning of any unfamiliar words, whose use is demanded by the plot of the story but has a low frequency, is given in parentheses upon its first appearance in the story. Words so explained are disregarded in the computation of the vocabulary density and the pupil is not expected to make any conscious effort to learn them. (pp. 326–327)

The description closely matches the advice in M. West (1926). Such texts could never be described as authentic, yet the text coverage criterion alone appears to have held the most sway and applied to authentic texts anyway.

But even the apparently sacrosanct text coverage figure of “one new word in fifty” (M. West, 1926, p. 27) is not the final word on the matter. As Holley noted, M. West put forth a variety of figures. In Bilingualism, M. West (1926/2003) asserted that texts could be read pleasurably with text coverage as low as 96.67%, one word in thirty, but that text coverage as high as 98.57%, one word in seventy, would be ideal:

According to this standard a boy studying a Primer will have to look up every thirtieth word in a dictionary, or 7 per page; and a boy studying a first reading-book will have to look up every fortieth, or 5 per page. At these standards we consider that the boy will just be able to derive pleasure and continuity of meaning from his reading; though a higher standard, e.g., 40-50 in the Primer and 60-70 in the First Reader would be very desirable. (p. 271)

The recommendation in Construction of Reading Material for Teaching a Foreign Language (M. West, 1927) should probably be considered authoritative of M. West’s position since it was written as guidance for constructing texts of a similar nature and incorporates lessons learned from the original research. The recommended coverage there, however, is unambiguously not one word in fifty. Rather, it is one word is sixty, or 98.33%:

The maximum density compatible with ease and enjoyment was originally fixed at 1 new in 50 running words. But in addition to new words there are new usages which also act as slight obstacles to ease and enjoyment. To allow for these individually would be very complicated, especially as they vary greatly in difficulty. Hence they are allowed for in mass by lowering the new word density to 1-60. (p. 15)

The figure is also repeated multiple times throughout the 32-page booklet:

“It will be recollected that in these reading-books there are sixty running words of text for every new word introduced.” (p. 11)
“But, since there is one new word for every sixty running words, in a unit of 600-700 running words there would be ten or twelve new words to earn;” (p. 17)

“The Frequency Vocabulary is also at hand: from this suitable new words are selected for introduction into the text at approximately sixty-word intervals.” (p. 24)

The mesmerising allure of the one in fifty figure is understandable because it comports with results such as those reported in Hu and Nation (2000):

The results suggest that for this kind of reading learners need to know at least 98% of the running words in the text. That is, the density of unknown words should be around one in fifty. This is the same figure arrived at by West (1955:21) and Hirsh and Nation (1992) simply by considering the amount of interruption to the flow of reading. (p. 423)

Ideally research should confirm or refine previous research, so it is tempting to promote a particular figure when it is in agreement with current results. But it should be abundantly clear that M. West did not advocate a single, idealised text coverage figure for general reading for pleasure. In fact, a range of values were promulgated, depending on the level of the reader. Additionally, the texts which M. West referred to were not typical, authentic texts. They were intentionally created learning materials. These facts stand in stark contrast to the modern interpretations and citations of M. West’s findings and ideas.

5.2.2 Carver (1994)

Carver’s more than 100 peer-reviewed articles primarily develop and support a model of reading known as Rauding Theory. This theory posits that basic reading consists of five primary reading processes, also known as “reading gears” (Carver, 2000, p. 24): scanning, skimming, rauding, learning, and memorising. Each of the reading gears entails the use of various hierarchical components or subprocesses. Gear 5, commonly known as scanning, requires only the “lexical accessing” (p. 24) component and, because of that fact, higher average reading speeds are attainable. Each subsequent Gear requires an additional component which consequentially slows down the average reading speed attainable, presumably because greater mental processing power is required for each Gear.

According to Carver, normal reading should be seen as rauding, or Gear 3. In Rauding Theory, reading speed and comprehension are inversely correlated so rauding represents a sort of happy medium between the two opposing goals. Carver’s model of
reading is compatible with, and subsumes, the Simple View of Reading (Carver, 1993) which bisects reading into two factors:

- decoding
- linguistic (or listening) comprehension (Hoover & Gough, 1990).

Rauding is a portmanteau formed from reading and auding (Carver, 2002) and shows that normal reading is no different from normal listening (auding). The Simple View of Reading has not been well received by some reading researchers (e.g. Hoffman, 2009), nor have Carver’s ideas escaped similar reactions (e.g. Pearson & Kamil, 1977). One reason why the Rauding Theory is controversial is because some of its claims appear to be overly optimistic. For example, Grabe (2004) describes the possibility of a perfect correlation ($R^2 = 1$) between measures of reader ability and reading comprehension as “startling in its assertion” (p. 49). In the main, the model has not had significant impact within contemporaneous reading research, but of course not all of Carver’s ideas can be dismissed outright. There certainly must be some truth to the idea that different modes of reading will result in variable reading speeds and degrees of comprehension. It is likely these kernels of truth which explain Carver’s continued influence in some corners, such as L2 reading research who have warmly embraced some of the studies.

Out of the numerous studies, only one explicitly focused on vocabulary: “Percentage of Unknown Vocabulary Words in Text as a Function of the Relative Difficulty of the text” (Carver, 1994). Van Zeeland and Schmitt (2013) note that it is probably the only L1 English study which compares the amount of unknown vocabulary in a text and reading difficulty. The details of Carver (1994) were discussed in detail in section 2.2 (page 18), but a brief recapitulation may be beneficial here.

Carver was sceptical of claims that free reading leads to vocabulary development in L1 English readers (e.g. Nagy & Anderson, 1984; Krashen, 1989) and posed the following question: “What percent of words in a passage are unknown when the passage is relatively easy, matched to ability level, or relatively hard?” (Carver, 1994, p. 415) To investigate these claims, participant reading ability was compared with a measure of text difficulty. The difference between the two measures determined whether the texts would be read with relative ease or difficulty. Participants underlined all unknown words in the passages which were then tallied and correlated with the relative easy or difficulty estimates. Carver found that participants did not know about 1% of the words when text difficulty was equally matched to participant ability. The findings, as summarised in the article’s abstract, make it very clear that the data collected in the study did not support vocabulary acquisition through reading:
These findings do not support the theory that free reading results in large vocabulary growth because free reading is likely to involve relatively easy material that contains few, if any, unknown words. Indeed, these findings call into question the practice of devoting large amounts of classroom time to free reading if the purpose is to increase vocabulary or reading level. (p. 413)

When it comes to text coverage studies in L1 English, this study is frequently cited but, as with M. West (1927), it seems to have been only narrowly read. It is frequently summarised as a text coverage study with findings analogous to Hu and Nation (2000) but conducted in L1 English. Hedgcock and Ferris (2009), for example, describes the outcome of the study thus:

In Carver’s L1 study, participants who successfully comprehended sample texts and found them easy to read knew 100% of the basic words in the text; participants who found the texts to be relatively difficult still knew 98% of the basic vocabulary items [...]. (p. 298)

On the face of it, this summary is apparently an accurate account of the main findings in Carver (1994): participants were asked to read a variety of texts which were estimated to be above and below their reading level which made the texts easy or difficult to read. The troubling part is but a single word: comprehended. The words comprehend and comprehension only occur five times, excluding the references section, within Carver’s (1994) approximately 10,000 word article. One of the occurrences is in a direct quotation. One would expect the term to be used more frequently if it were a central concept in the findings of the study. In actual fact, Carver did not measure the comprehension of the participants in the study; they simply underlined words that they did not know when reading a variety of texts.

Other citations omit the relative ease and difficulty aspect completely. Rodgers (2013) characterises the study as a text coverage prescription for reading comprehension: “Carver (1994) proposed an even more conservative estimate of 98% to 99% coverage as being necessary for comprehension.” (p. 106) It is difficult to see how this conclusion can be drawn when comprehension was not measured nor does the term appear anywhere in the abstract, results, or discussion sections of the article. Yet such unjustified conclusions are far from uncommon.

Although the study featured L1 English readers, Carlo et al. (2004) not only cited the study in support of a supplemental vocabulary programme for L2 English learners, but also transformed the conclusions into a prescription for both reading comprehension
5.2. SUMMARIES OF PREVIOUS RESEARCH

and vocabulary acquisition: “Relying on incidental vocabulary learning is even more problematic for ELLs than for their English-only (EO) counterparts. [...] Reading texts in which more than 2% of the words are unfamiliar blocks comprehension and novel word learning (Carver, 1994).” (p. 191) These conclusions would be somewhat justified had the study been experimental rather than observational.

Webb (2010) incorrectly suggests that Carver (1994) was experimental, putting it on the same footing as Hu and Nation (2000): “Studies that have manipulated the coverage of a text have found that increasing coverage above 90% is likely to improve comprehension (Carver, 1994; Bonk, 2000; Hu & Nation, 2000).” (p. 504) Such a claim is entirely understandable given that Hu and Nation also asserted similarity between their study and Carver (1994): “The findings agree closely with a study of native speakers by Carver (1994)” (p. 422). But claiming similarity or equivalence is unjustified. The study was not experimental in even the loosest sense of the word. Only experimental studies can make claims of causality.

While some citations avoid adding reading comprehension to the study, other citations fall afoul of Hume’s law by converting is to ought. Although Carver sought to measure the reality of unknown word encounters when reading material matched to the reader, I. S. P. Nation (2001) characterised the observation (i.e. is) as a recommendation (i.e. ought) for learning vocabulary through meaning-focused input: “Carver (1994) argues that for native speakers [text coverage] should be around 99%.” (p. 388)

To their credit, van Zeeland and Schmitt caution that “it should be emphasized that Carver relates word knowledge to text difficulty and not to text comprehension.” (p. 458) Compared with other citations which ascribe a strong connection with reading comprehension, van Zeeland and Schmitt seem to have taken up the opposite position. But it is easy, and common, to argue that in the case of Carver (1994), a relationship between reading comprehension and difficulty does exist and is referenced in the article. Carver clearly notes that

> [t]he functional relationship between the relative difficulty of a passage and the degree to which a passage can be comprehended was studied by Carver (1990a). [...] [T]he accuracy of comprehension while engaged in normal reading, called rauding, was predictable from the relative difficulty of the material. (Carver, 1994, p. 414)

Additionally, the measure of reader ability, the Accuracy Levels Test, is an assessment which predicts 75% comprehension on reading material which is matched in difficulty to the reader (see p. 417).
The truth of the matter is ambiguous at best. Carver’s Rauding Theory predicts that rauding occurs at the optimal rate of comprehension but the study’s research question and conclusions are obviously unrelated to reading comprehension. It seems that some researchers prefer to note that the participants’ predicted reading difficulty levels were positively correlated with the percentage of unknown words in the text, which can then be combined with previous research (Carver, 1990) finding a negative correlation between perceived difficulty ratings and reading comprehension.

To draw this inference, as Carver does, is to essentially engage in syllogistic reasoning. The argument is strikingly similar to a hypothetical syllogism which takes the form \( P \rightarrow Q, Q \rightarrow R \vdash P \rightarrow R \) and is valid. However, correlations, unlike conditional statements, are not transitive. The true form of the inference is \( \text{corr}(\alpha, \gamma), \text{corr}(\beta, \gamma) \vdash \text{corr}(\alpha, \beta) \). But it is entirely possible for two variables, \( \alpha \) and \( \beta \), to both be correlated with a third variable, \( \gamma \), whilst exhibiting no correlation between \( \alpha \) and \( \beta \) whatsoever. Only when \( r^2 > .50 \) for both \( \text{corr}(\alpha, \gamma) \) and \( \text{corr}(\beta, \gamma) \) must \( \text{corr}(\alpha, \beta) \) be greater than zero. (Stanley & Wang, 1969). Carver provides no correlation statistics so it is impossible to determine whether the inference is justified. In other words, unless the correlations between text coverage and relative difficulty in Carver (1994) and between relative difficulty and reading comprehension in Carver (1990) were both > .707, then it is impossible to conclude that there is any correlation between text coverage and reading comprehension. The correlations for both studies must have been > .707 in order for the relationship between text coverage and reading comprehension to be a mathematical certainty.

However, even if there were no confusion in the relationship between Carver (1994) and Carver (1990), no significant problems in conflating correlation with causation, nor suspicion in the recharacterisation of the results detailed here, there are severe methodological weaknesses in both Carver (1994) and Carver (1990) laid out in subsection 5.3.3 (page 131) and subsection 5.3.1 (page 124) which would preclude serious consideration of this study as the foundation for further research.

### 5.2.3 Hu and Nation (2000)

The original study which prompted the replication and expansion studies detailed in chapter 3 (page 59) is not immune either to overly optimistic interpretations of the results. The study was one of the few that employed an experimental design and thus claims of causality are warranted unlike the correlation-based studies. But, as has been shown previously, adding causality is not the only problem apparent in summaries of previous research.
Salah (2008), for example, seems to have found in the study not only an ideal text coverage but also the vocabulary size necessary to reach that coverage:

The study of Hu & Nation (2000) concluded that readers need at least a vocabulary of 5,000 word families to cover approximately 98% of the running words in a novel and to achieve adequate unassisted comprehension. Few L2 readers were able to comprehend well with around 90%-95% vocabulary knowledge in a given text. (p. 17)

Perhaps this is a misattribution, though. The 5000 figure in the first sentence appears to be referring to Hirsh and Nation (1992), which is also cited in Salah (2008) but the second half of that sentence is clearly a paraphrase of a statement in Hu and Nation (2000). The former was not an experimental study so reading comprehension of participants was not measured. But no matter the provenance of the figure, it certainly is not from Hu and Nation (2000) and it exemplifies the general problem in summarising and integrating previous research into a theoretical framework from which future research will be hung.

Other citations raise or lower Hu and Nation’s (2000) text coverage recommendation of 98% for unspecified reasons. N. Schmitt et al. (2011), for example, raise the figure by one percentage point: “However, Hu and Nation (2000) reported that their participants needed to know 98%–99% of the words in texts before adequate comprehension was possible.” (p. 26) It is not clear whether N. Schmitt et al. noticed the calculation error detailed in section 4.4 (page 108) because there is no explicit mention of it in their article. But changing the text coverage recommendation is common. Gardner (2004), for example, cites Hu and Nation (2000) as advocating a lower text coverage figure:

Interestingly, several studies have also concluded that the threshold of vocabulary knowledge for adequate reading comprehension is around 95 per cent of the words in a given text (Hu and Nation 2000; Laufer 1989, 1997; Liu and Nation 1985; Wixson and Lipson 1991). (p. 4)

Despite the caveat from the authors that the study’s results would only apply to “reasonably easy [fiction] text[s]” (p. 422) and that other genres would likely require even higher text coverage figures, many studies which cite this research fail to heed the warning. Tanaka and Stapleton (2007), for example, understood the results to mean that 98% text coverage is needed “for the purpose of language growth” (p. 119) in an extensive reading program. Although reading for language growth is discussed in Hu and Nation’s conclusion, it is in the context of distinguishing three types of reading, intensive, extensive, and reading for language growth, the last of which most closely
matches the experience of the participants in the study. But Hu and Nation also clearly state that the connection between 98% text coverage and vocabulary acquisition is not clear by asking researchers to investigate whether or not “most unknown words are easily guessed at the 98% coverage level” (p. 423). Similarly, Miekley (2005) created a checklist to be used when evaluating L2 English textbooks which incorporates the 98% text coverage figure as one of the criteria despite the obvious differences in textual features between textbooks and fiction texts.

The irony is that text coverage studies in general seem to be concerned with refining a particular figure to a great degree of accuracy, but researchers do not seem to apply the same standard to interpreting and situating previous text coverage research. The lax manner in which the methodologies, results, and conclusions are recast and summarised virtually ensures that the relationship between text coverage and reading comprehension, if any exists at all, will become less, rather than more, clear over time.

5.3 Methodologies and operationalisation of terms

There exists a surprising amount of diversity in the way terms are operationalised in text coverage research. Likewise, methodologies for researching the relationship between text coverage and reading comprehension differ substantially without a clear justification for doing so. This lack of precision and occasional conflation of terms is one of the reasons why the relationship is still not very clear despite studies spanning decades.

5.3.1 Reading comprehension

Reading comprehension has been a topic of formal research for more than a century. Yet despite a long tradition and multiple approaches, the concept today remains only slightly clearer than it was when the earliest researchers attempted to operationalise it. Nevertheless, the lack of consensus on what exactly reading comprehension is, and how to accurately assess it, has not deterred many from inventing and promoting seemingly countless reading programs and interventions purportedly designed to improve it, often with the claim that the techniques employed are based on the latest research in reading comprehension.

The issue is further challenged by the myriad approaches with which reading comprehension can be studied, each bringing its own set of questions and interests. As an example, consider the following summary of findings from previous research as summarised in Laufer and Ravenhorst-Kalovski (2010): "Hu and Nation (2000) [...] used two
5.3. METHODOLOGIES AND OPERATIONALISATION OF TERMS

comprehension tests and defined ‘adequate’ comprehension [...] If we average out the two scores, we receive 71%.” (p. 17) Combining the two measures of comprehension in this way is understandable; Hu and Nation draw little distinction between the two other than noting that the first test, an MCQ test, contains more items targeting implicit propositions while the second test, a written cued recall test, primarily targets explicit propositions. It is possible, though, that the two tests measure slightly different constructs. For example, compared with the MCQ test, the written cued recall test is basically a proposition-by-proposition retelling of the story, focusing almost exclusively on explicit propositions. Another difference is that the ratio of propositions in the text which appear in the written recall test is much higher than that of the MC test.

By simply combining the two measures, Laufer and Ravenhorst-Kalovski (2010) make the tacit assertion that the two reading comprehension are interchangeable. But are they? N. Schmitt et al. (2011) take the same position and directly compare the relative performance on different measures of reading comprehension:

Hu and Nation (2000) found that even at 80% coverage, their learners were able to score about 20% on the written recall test, and 43% on the MC comprehension test. At 90% coverage, this improved to 41% WR and 68% MC. In our study, the learners could answer 50% of the comprehension items correctly at 90% vocabulary coverage. This suggests that although comprehension may not be easy when there is more than 1 unknown word in 10, learners can still achieve substantial comprehension. (N. Schmitt et al., 2011, p. 35)

Despite the tests measuring different aspects of comprehension, in different ways, N. Schmitt et al. confirm the tacit assumption that all reading comprehension measures can be directly compared.

N. Schmitt et al. (2011) do not dwell on the fact that participants were “free to return to the texts as much as they wished when answering the comprehension items” (p. 33) while participants on Hu and Nation (2000) knew in advance that the text would not be available. The consistency with which the two sets of participants would have approached these very different tasks can not be assumed. One measure involves the ability to find information in a text and transfer it to another form while the other involves reading to remember followed by tests of unknown format. Generalising the idea further, one might wonder whether any and all measures of reading comprehension used in text coverage studies are equivalent to the extent that they can be directly compared.
Reading comprehension results from text coverage studies are rarely the same but no justification has been given for assuming that the results are equivalent.

Cutting and Scarborough (2006) addressed this problem head on by comparing three measures of reading comprehension which are widely used in L1 English reading research. The motivation for such a study succinctly encapsulates the problem of assuming that different tests of reading comprehension can be directly compared:

> If different skill sets are more important to performance on some reading comprehension tests than others, conclusions about the nature of comprehension (and comprehension difficulties) will be specific to the test, leading to the kinds of mixed findings that have been reported about the relative contributions of bottom-up skills and various cognitive and linguistic abilities. (pp. 281–282)

When Cutting and Scarborough regressed the reading comprehension measures on various factors thought to contribute to reading comprehension, the tests were found to vary in the construct they purport to measure. Before comparing reading comprehension measures from disparate studies, researchers should account for this difference.

Recalling that the theoretical foundation of text coverage studies is the notion of a lexical threshold which limits the availability of generalised reading skills, it is certainly not a stretch to suspect that small-scale ad-hoc tests of reading comprehension of individual texts like the ones found in text coverage research can not be assumed to measure the same construct of reading comprehension. Although there is likely overlap in theses studies, the degree to which they do is unknown. A pessimistic view of the situation might conclude that text coverage studies have taken N. B. Smith’s (1963) quip on reading comprehension to heart: “‘Comprehension’ is just a big blanket term that covers a whole area of thought-getting processes in reading.” (p. 257) If that is the case, then it is understandable why little attention has been paid to clearly stating and operationalising the second half of a two part model and instead attention has primarily been focused on the first half.

But the problem is not unique to text coverage studies. When it comes to L2 reading research in general, there seems to be little consensus as to what reading comprehension actually is and how to measure it, yet the term is often used in a way which implies that universal agreement exists so it is not worth further elaboration. Bernhardt (2005), for example, in an article which introduces a model of second language reading, men-

---

1Details on the theoretical underpinnings and evolution of the text coverage model of reading comprehension can be found in section 5.4 (page 140).
tions the term *comprehension* only seven times in the ≈ 4400 word article and never clearly defines the term other than noting it consists of “different elements and influences” (p. 138) which cannot be explained by a simple combination of grammar and vocabulary. Perhaps the concept is self-evident?

The truth is that the vast majority of tests of L2 reading comprehension are “largely atheoretical” (Daneman, 1982, p. 332) and consequentially, inconsistent in clearly defining the construct the tests measured. One early text coverage study described the reading comprehension test as one in which “the subjects were required to show comprehension of general ideas, to elicit and interpret factual information, supporting material, to draw inferences. In other words, they were performing tasks similar to those any reader of academic texts would be confronted with.” (Laufer, 1989) With this description, it is difficult to know what construct was measured and how the measurement could be replicated. Bormuth (1971) asserts that small scale tests created by teachers are almost certainly too diverse in construct validity and item difficulty to be reliable.

One particularly egregious study merits noting here to serve as an exclamation mark on this issue. While Carver (1990) was neither a study of L2 reading nor a text coverage study, it is cited in Carver (1994) which itself is often misattributed in L2 text coverage studies as perhaps the L1 lone text coverage study.² The former study found a linear relationship between reading comprehension and relative difficulty while the latter study found a linear relationship between relative difficulty and the percentage of unknown words in a text. These two correlations make it tempting to bridge the gap and conclude that text coverage is correlated with reading comprehension, but a careful examination of the measurement method employed in Carver (1990) should make even the most sympathetic researcher pause. To assess reading comprehension “students were asked to decide what percent of each passage they comprehended after they had read it, and this judgement constituted an estimate of the accuracy of comprehension.” (p. 405) The primary problem with this approach is the inevitable influence of illusory superiority, a well-known phenomenon also known as by the eponym *Dunning–Kruger effect* (Kruger & Dunning, 1999). It is also known more colloquially, and humorously, as the *Lake Wobegon effect* (Koretz, 1988) which refers to a fictional place where “all the children are above average” (Keillor, 1993). Essentially, illusory superiority describes the process where subjects’ perception of their performance or ability bears little resemblance to their measured performance or ability.

²Details on the ways Carver (1994) is incorrectly summarised in L2 text coverage studies can be found in subsection 5.2.2 (page 118).
<table>
<thead>
<tr>
<th>source</th>
<th>explicit</th>
<th>sequence</th>
<th>causal relations</th>
<th>main idea</th>
<th>inference</th>
<th>author purpose</th>
<th>author style</th>
<th>self assessed</th>
<th>MCQ answer</th>
<th>short text</th>
<th>text available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holley (1973)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>?</td>
<td>?</td>
<td>×</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Laufer (1989)</td>
<td>×</td>
<td>?</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Carver (1990)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>?</td>
<td>×</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Hu and Nation (2000)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Laufer and Ravenhorst-Kalovski (2010)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>N. Schmitt, Jiang, and Grabe (2011)</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>van Zeeland and Schmitt (2013)</td>
<td>×</td>
<td>?</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

Table 5.1: Comparison of the operationalisation of reading comprehension.
While Carver (1990) has rightly been ignored by most reading researchers, the scepticism such a methodology engenders should not be reserved for such obviously flawed measures. Even those which appear to be high in construct validity should be rigorously examined. Perhaps the most direct method of understanding the nature of the problem of comparing these measures is to contrast the aspects of reading comprehension which are actually measured, and the methods for doing so, in several key studies. The characteristics of the reading comprehension measures used in a selection of studies which feature prominently in text coverage research can be found in Table 5.1 (page 128). Given the difficulty in measuring the same construct by a small number of standardised tests of reading comprehension (Cutting & Scarborough, 2006), the variety of aspects of reading comprehension and the formats used in typical text coverage studies should raise serious suspicions about the constructs measured and the soundness of directly comparing them.

### 5.3.2 Adequate

The operationalisation of the term *adequate* comprehension is acknowledged as being a somewhat arbitrary decision (Hu & Nation, 2000, p. 414), but a reasonable criterion is used to establish a minimal level which should be considered adequate. That criterion is the level at which most of the participants in the 100% text coverage condition can be classified as having attained adequate comprehension. In other words, if one’s performance on the reading comprehension measure is at or above the threshold, it would be safe to assume that the reader was performing at least as well as a typical reader would if all of the words were known.

This threshold is not only intuitively reasonable, it is easy to calculate: setting the threshold for adequate comprehension as the median score of the participants in the 100% text coverage condition guarantees, by definition, that the majority will be at or above the threshold. This is a reasonable approach to take, but Hu and Nation’s (2000) implementation is inconsistent. For the MCQ reading comprehension measure, the threshold for adequate comprehension is the nearest integer (12) below the mean (12.24). For the written cued recall, however, the bar is set much lower at “around 70” (Hu & Nation, 2000, p. 416) from a mean score of 77.17. The justification for both is that the proportion of scores above these thresholds are greater than those below it. But the cued recall scores are aggregated into bins of 10 points each without any justification for doing so. So although 77.17 was nearer the top of the 70–79 bin, the threshold was set at the bottom.

Without going so far as to accuse Hu and Nation of exercising too many researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011), one cannot help but wonder if this threshold was set specifically to match the 98% text coverage figure from the MCQ test. If the threshold had been set at the midpoint or top of the bin, the text coverage
required to attain that level of reading comprehension would be 100% rather than having the two measures perfectly agree.

Although there are no data to support any nefarious intentions in selecting that particular threshold for adequate comprehension, other researchers have taken a less than principled approach to operationalising this term. Laufer and Ravenhorst-Kalovski (2010) converted the absolute criteria for adequate comprehension on the two measures in Hu and Nation (2000) to relative criteria and averaged them to compare with relative criteria in other studies. The rational for taking an average of multiple measurements is that error decreases with aggregation. However, this rule only applies to sampling error. If Hu and Nation’s (2000) two figures for adequate comprehension are considered as repeated measures of the adequate construct, then the conclusion should be that true threshold for adequate reading comprehension based on the 14-question MCQ test is $\approx 10$. Such a conclusion would then shift the recommended text coverage figure from $99\%$\textsuperscript{3} to $93\%$ based on the MCQ test and $100\%$ based on the cued recall test. The threshold for the cued recall test would be so high that $> 95\%$ of all participants in the study would not have achieved adequate comprehension.

The cut-off point for adequate comprehension is, of course, arbitrary, but Laufer and Ravenhorst-Kalovski (2010) are correct in focusing on the relative performance on reading comprehension measures to facilitate initial comparisons. For example, Bormuth (1971) investigated both 75% and 90% reading comprehension test criterion which were traditionally used but did not appear to enjoy the support of data. The Lexile® Framework (MetaMetrics, Inc., 2014) matches readers to texts by predicting a “75% comprehension rate” (para. 4) which is described as “comprehend[ing] approximately 75% of a book” (para. 4). Conceding that there may be a range of criteria which could be considered as adequate, MetaMetrics, Inc. (2007) puts forth a strong case for the 75% criterion as adequate performance on well-designed measures of reading comprehension, noting that empirical data support it as a balance between too difficult and too easy, particularly for young readers.

A non-random sample of studies and the criteria adopted to operationalise adequate comprehension can be found in Table 5.2 (page 131). The lack of not only an agreement on what constitutes reading comprehension, but also a criterion for the minimum acceptable reading comprehension characteristics makes it difficult, if not impossible, to directly compare the limited number of text coverage studies and even with the wider body of research on readability studies. The question of exactly what reading comprehension entails and how best to operationalise it will not be resolved in the near future,

\textsuperscript{3}The reported figure of 98% is due to a miscalculation explained in section 4.4 (page 108).
5.3. METHODOLOGIES AND OPERATIONALISATION OF TERMS

Table 5.2: Comparison of reading comprehension thresholds.

<table>
<thead>
<tr>
<th>source</th>
<th>threshold</th>
<th>measure</th>
<th>reading type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thorndike (1914)</td>
<td>80%</td>
<td>Scale Alpha¹</td>
<td>reading proficiency test</td>
</tr>
<tr>
<td>M. West (1926/2003)</td>
<td>75%</td>
<td>underline text</td>
<td>reading proficiency test</td>
</tr>
<tr>
<td>Bormuth (1969)</td>
<td>50%</td>
<td>cloze</td>
<td>assisted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80%</td>
<td>unassisted pleasurable</td>
</tr>
<tr>
<td>MetaMetrics, Inc. (2014)</td>
<td>75% cloze²</td>
<td></td>
<td>&quot;targeted&quot; reading</td>
</tr>
<tr>
<td>Laufer (1989)</td>
<td>55%</td>
<td>comprehension test</td>
<td>high-stakes exam</td>
</tr>
<tr>
<td>Hu and Nation (2000)</td>
<td>85.7% MCQ</td>
<td></td>
<td>for pleasure³</td>
</tr>
<tr>
<td></td>
<td>56.5% cued recall</td>
<td></td>
<td>for pleasure³</td>
</tr>
</tbody>
</table>

¹ Scale Alpha is considered to be the first standardised silent reading test (Tuinman, 1971) consisting of short-answer questions about paragraphs of increasing difficulty.
² Original development used the cloze test but the Lexile® Framework can be calibrated to any reading task but the performance forecast is based on the same threshold.
³ See critique of the validity of this claim in subsection 2.7.3, page 48.

but some interim standard must be adopted before text coverage studies, or any studies of reading comprehension, can be compared directly.

5.3.3 Text coverage determination

There are three ways to determine the text coverage experienced by a reader:

- self-reporting of unknown words in the text,
- calculating the overlap between the vocabulary size of the reader with the lexical profile of the text,
- predicting that the reader knows all words in the text except for any surreptitiously placed pseudowords.

All three methods have been used in text coverage studies but little, if any, consideration has been given to the relative benefits, consequences, and comparability of the methods.

5.3.3.1 Self-reporting

Self-reporting has been used in both L1 (e.g. Carver, 1994) and L2 (e.g. Laufer, 1989) text coverage studies. Typically participants are asked to underline or otherwise indicate only words that are not known. One problem with this method is that it relies
somewhat on the honesty of the participants. Reporting accuracy can be cross-checked through a task which demonstrates both little to no knowledge of the indicated words as well as some knowledge of the remaining words which the participant claims to know. Laufer (1989) sampled 40 words to use for verification and adjusted self-reported figures based on the verification results. There will always be some uncertainty in the actual text coverage unless all words are verified, though. It is possible, but not practical, to verify all of the words in a text but if all words are tested, then self-reporting becomes superfluous. Therefore a balance must always be struck between the acceptable level of uncertainty in the measurement and the practicality of measurement.

A more general, and less reliable method, of verification is to include some very low frequency or pseudowords which can be safely assumed to be unknown to all participants. Failure to indicate these words identifies participants who were not honestly reporting word knowledge. Carver (1994) used this method in training materials and then simply excluded participants who failed to indicate the low frequency words as unknown. This can create a biased sample, though. Participants excluded due to their inability to indicate unknown words are very unlikely to be randomly distributed among the group. They might have an overly inflated sense of confidence, be less averse to taking risks, or possess some other characteristic which may influence the results of the study. A better approach would be to offer further training to this subset of participants. Even if that occurred, however, the method relies on the assumption that the strategy for determining unknown words remained the same for every participant throughout the data collection period.

Another issue with self-reporting to determine text coverage is the perennial problem of operationalising the degree of word knowledge a participant must possess before a word can be claimed as known, or in this case, how little word knowledge may be possessed and still claim that the word is unknown. Carver (1994) instructed adult participants to indicate a word was unknown if the participant’s knowledge was at Stage 1, *never seen before*, or Stage 2, *recognised but no knowledge of meaning*, of Dale’s (1965) scale of word knowledge. Laufer’s (1989) criteria was slightly different; words which were unknown to the reader but understandable from the context were to be regraded as known. It is not clear how participants were to verify the accuracy of their newly acquired word knowledge, however.

Finally, it is not clear how accurately participants can assess their own ability especially for borderline cases. Partial knowledge which may or may not be correct may result in a claim of knowledge for some confident participants but indicated as unknown for others. One solution, implemented by N. Schmitt et al. (2011), is to use a yes/no test
constructed from all, or at least a large portion, of the words in the text. By including some pseudowords in the test, researchers can verify that participants are honestly indicating knowledge of the words, but only at the level of form recognition; knowledge of the meaning associated with the form can only be inferred.

5.3.3.2 Overlap between vocabulary size and cumulative frequency

To measure knowledge of a word’s meaning, a test such as the VST can be used to estimate participants’ vocabulary size which can then be used to estimate text coverage. In many text coverage studies, frequency lists derived from the BNC are used and the tacit assumption is made that a frequency list from a large, balanced corpus such as the BNC can be a reasonably accurate model of an L2 English learner’s mental lexicon. This assumption can clearly be seen in Laufer and Ravenhorst-Kalovski (2010): “[...] we replace the BNC lists with learners’ vocabulary size. If, for example, 5,000 words cover 94% of a text, then learners with a knowledge of 5,000 words can understand a similar percentage of this text.” (p. 23)

To elaborate, the assumption is that the 5,000 words in this hypothetical learner’s mental lexicon can be reasonably modelled by the first 5,000 words, in rank order, according to their frequency of occurrence in the BNC and thus the two are interchangeable. Hirsh and Nation (1992) explicitly stated that these types of calculation rely on “the assumption that vocabulary is learned in relation to its frequency of occurrence” (p. 692). This assumption can be seen in action in I. S. P. Nation (2006) which investigates text coverage provided by the BNC word lists over a variety of text genres to determine vocabulary learning goals: “With a vocabulary of 4,000 word-families and assuming that proper nouns are easily understood, 95.06% of the tokens would be familiar.” (p. 70) An example of the increasing text coverage provided by each list of 1000 can be found in Table 5.3 (page 134) which is adapted from tables found in I. S. P. Nation (2006). This assumption forms the theoretical foundation for a large number of corpus-driven text coverage studies, more than a dozen of which are catalogued in Webb and Macalister (2013) in support of the claim that the methodology for determining the vocabulary size needed to reach a particular text coverage is “well established” (p. 303).

Unfortunately, the assumption is not a safe one to make. To see why, consider how Fan (2013) defines text coverage formally:

\[
T_c = \frac{\sum_{i=1}^{S} F_{wi}}{N}
\]
Table 5.3: Tokens and cumulative percentage coverage figures for *Lady Chatterley’s Lover* by 1,000 word-families from the BNC (adapted from I. S. P. Nation, 2006).

<table>
<thead>
<tr>
<th>Word List Tokens (1,000)</th>
<th>Coverage including proper nouns (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80.99</td>
</tr>
<tr>
<td>2</td>
<td>7.21</td>
</tr>
<tr>
<td>3</td>
<td>3.14</td>
</tr>
<tr>
<td>4</td>
<td>1.78</td>
</tr>
</tbody>
</table>

where $T_c$ is the text coverage provided by a list of words ($W_{list}$) over a text or collection of texts ($T$) containing $N$ tokens, $S$ is the number of lemmas in $W_{list}$, $F_{wi}$ is the frequency of occurrence of the $i^{th}$ inflectional lemma form ($w_i$) in $T$ which is $\geq 0$. As a simplified example, if a word list ($W_{list}$) consisted of just three lemmas with two inflections each, four of which occurred one time each in a text, $T$, consisting of $N = 5$ unique tokens, the text coverage would be

$$T_c = \frac{0 + 0 + 1 + 1 + 1 + 1}{5} = 0.8$$

or 80% text coverage. It should be clear from the formal description and the example that the accuracy of the text coverage determination delicately hinges on how accurately the contents of $W_{list}$ can be known. When $W_{list}$ is derived from a corpus for a modelling study such as Fan (2013), the accuracy is perfect. But when $W_{list}$ is a reader’s mental lexicon and a word list is used as a proxy, the accuracy lies in the degree of overlap between the two. Unfortunately, the overlap is not perfect.

Perhaps the assumption of interchangeability is understandable when the proportion of a text represented by each subsequent list of 1000 word families, as found in Table 5.3 (page 134) drops so dramatically, it appears that the amount of error in making the assumption would be small. A graphical illustration, Figure 5.1 (page 135), of the text used in Hu and Nation (2000) and this thesis and a hypothetical reader suggests that the error should be inconsequential. In the plot, the number of unique word families found in *The Escaped Madman* from each of the first eight 1000 word family lists appear along with the number of unique word families in each of the lists known by a hypothetical reader who has a vocabulary size of 3000 word families. According to a frequency profile of *The Escaped Lunatic*, the first three 1000 word family BNC lists provide 97.93% text coverage, so according to the logic in Laufer and Ravenhorst-Kalovski (2010) and I. S. P. Nation (2006), a vocabulary size of 3000 word families should give the same coverage.
5.3. METHODOLOGIES AND OPERATIONALISATION OF TERMS

Figure 5.1: Comparison of total word families in terms of proportion of word frequency bands in *The Escaped Lunatic* and a reader’s mental lexicon with a vocabulary size of 3,000 word families.

Figure 5.2: Overlap of a hypothetical reader’s 3,000 word family mental lexicon (grey squares) and word families in *The Escaped Lunatic* which would probably be known (solid squares) and unknown (open red circles) to the reader. A frequency profile indicates 98% text coverage at the 3k level (dashed line).
But the plot is deceiving. Because the $y$ axis represents the total number of word families, not the actual word families themselves, the frequency profile of the text appears to be contained, in the lower left-hand corner of the plot, within the vocabulary size of the reader. What the viewer of the plot would be forgiven for missing is that the upper right-hand section represents unknown word families.

This phenomenon is noted in van Zeeland and Schmitt (2013) but the severity of the problem is underestimated. Van Zeeland and Schmitt exclude participants who fail to achieve near perfect performance on the 2000 level of the VLT because the experimental design, as with Hu and Nation (2000) and discussed in detail in subsubsection 5.3.3.3 (page 138), requires complete knowledge of the high frequency words in the text. By setting such a strict criterion for inclusion in the statistical analyses, van Zeeland and Schmitt hoped to minimise the potential for error in estimating text coverage: “Having one mistake on the whole VLT test suggests a participant could be unfamiliar with up to a maximum of 3.3 per cent of the words in the 2,000 level (i.e. one error from 30 target words)” (p. 468) Ignoring the overly optimistic error ceiling of 3.3% which assumes no measurement error in the VLT, the degree of effect on text coverage this lack of knowledge might have is contingent upon the number of times the unknown words are used in the text. Unknown words exhibiting a higher than average type-token ratio have a much greater effect than the predicted 3.3% claimed by van Zeeland and Schmitt. For example, by using the mean text length (475.5) of the four experiment texts in van Zeeland and Schmitt (2013) and the mean type-token ratio (38.25), the mean types per text can be calculated:

$$t_{\text{length}} \times \frac{r_t}{100} = 475.5 \times \frac{38.25}{100} = 181.88.$$  

On average, each of those $\approx 182$ types occurs 2.61 times. But word frequencies are discrete, so the effect of an unknown token which occurs more often than 2.61 times will bring the text coverage lower than the $100 - 3.3 = 96.7\%$ expected by van Zeeland and Schmitt.

To see how this works, a term for unknown vocabulary can be added to Fan’s (2013) text coverage formula:

$$T_c = \frac{\sum_{i=1}^{S} F_{wi} - \sum_{i=1}^{U} F_{wi}}{N}$$

where $U$ is the number of unknown lemmas in the text. The assumption is that all of the words are known except for 3.3% predicted by one incorrect response on the VLT,
so the rounded mean text length can be substituted for both $N$ and $\sum_{i=1}^{S} F_{wi}$:

$$T_c = \frac{476 - \sum_{i=1}^{U} F_{wi}}{476}.$$ 

The estimated number of unknown types is 3.3% of the $\approx 182$ mean unique types per text, or $U = 6$, each occurring 3 times, or the integer closest to the mean occurrences per type. The result is

$$T_c = \frac{476 - (6 \times 3)}{476} = 0.9622$$

or 96.2% text coverage. Unknown tokens exhibiting an even greater frequency in the text would bring the text coverage even lower.

The small difference may seem insignificant, but in the tails of power distributions, such as the Zipf distribution, a small change in one axis is a large change in the other. Van Zeeland and Schmitt (2013) give an example of this phenomenon, noting than only 2000 to 3000 word families would need to be known for 95% text coverage, but more than double that, from 6000 to 7000 word families, for 98% text coverage, a change of just three percentage points. To make matters worse for van Zeeland and Schmitt’s (2013) assumption, the unit of measure used for the participants’ vocabulary size is word families but the texts are described in terms of type-token ratio, not family-token ratio, so the actual text coverage would be even lower because there are likely to be more than one type per family. The study investigated the effects of 100%, 98%, 95%, and 90% text coverage levels, but the error in determining text coverage, $\approx 4\%$, for more than a third of the participants based on the overlap between vocabulary size and frequency profile is double the distance between the first two conditions.

The attempt by van Zeeland and Schmitt (2013) to limit the error inherent in this method of estimating text coverage is commendable, but the severity and nature of the problem could be more explicit. As just described, the impact of unknown words which occur more frequently in a text will strongly affect the text coverage. Plotting the overlap between the two as in Figure 5.1 (page 135), it appears that lower frequency words which, by their nature, should occur less frequently in a text would dominate the unknown category. When the same figures are plotted as points randomly distributed throughout their respective 1000 word lists as in Figure 5.2 (page 135), it becomes apparent that unknown low frequency words actually represent a minority of the total unknown words represented by open red circles. The dashed line indicates the end of the 3000 word list, so knowledge of all of the word families to the left of this line would provide $\approx 98\%$ text coverage. A 3000 word family mental lexicon, however, contains
only about 2100 of the words in those lists, represented by solid grey squares, which leads to a surprisingly high number of high frequency unknown words. The overwhelming impact that just a few unknown high frequency words can have on text coverage is significant. Even studies with strict vocabulary knowledge criteria such as N. Schmitt et al. (2011), where only around 15% of the participants failed to attain flawless performance on a test of the first 1000 words, suddenly become suspect. Lack of knowledge of even a single high frequency word can clearly have a much greater impact on the actual text coverage experienced by the reader than the estimated text coverage based on the overlap methodology. This casts serious doubts on the assertion that a similarly sized frequency list can be seamlessly substituted for an estimate of a reader’s vocabulary size (Laufer, 2013).

5.3.3.3 Pseudowords

The last method of measuring text coverage is a special case of using the overlap between vocabulary size and cumulative frequency from a lexical profile. The goal is to use the overlap to verify that the text coverage is close to 100% and then replace some words with pseudowords to ensure that the text coverage very closely follows the proportion of pseudowords. The method has been used in a variety of text coverage studies (e.g. Hu & Nation, 2000; Kurnia, 2003, and the experiment detailed in chapter 3). Because it is a special case, this method is vulnerable to not only the concerns outlined in subsubsection 5.3.3.2 (page 133) but also another problem: selection bias.

The most common procedure for selecting words to replace with pseudowords is to rank the types, lemmas, or word families in increasing order of frequency and then select the requisite number, beginning at the top of the list, needed to achieve the desired text coverage. This results in a text with a lexical frequency profile which does not have a long tail; it will have an abrupt precipice. Perhaps a better, more ecologically valid, approach would be to replace words from a range of frequencies, with less frequent words more likely to be selected, but not exclusively.

This unfortunate artificial construction is concerning when the distribution of keywords in a naturally occurring text is considered. Keywords are words which are statistically more likely to carry meaning in a text and they are usually lower frequency words which naturally show up in that long tail which means they are more likely to be replaced by pseudowords using this method.

The most common method for determining keywords in a text within the field of information retrieval is known as TF×IDF (Robertson, 2004). It is the product of two text metrics: term frequency (TF) and inverse document frequency (IDF). TF is simply
the number of times a term, whatever the unit of measure, is found in a text and can be defined formally as

$$\text{tf}(t) = f(t, d)$$

where $f$ is the frequency of term $t$ in document $d$ although variants do exist to limit the influence of long documents. Likewise, many variants exist for determining IDF but the original concept (Spärck Jones, 1972) is described formally by Robertson (2004) as

$$\text{idf}(t_i) = \log \frac{N}{n_i}$$

where $n_i$ is the number of texts in a collection of $N$ documents in which term $t_i$ appears. To avoid division by zero when the term is not found anywhere in the collection, $1 + n_i$ is often used for the denominator. TF×IDF does not indicate a threshold for significance; it simply weights terms in order of importance or the proportion of meaning carried which is sufficient for most information retrieval. Tests of statistical significance can be used to determine a finite set of keywords when needed (Scott, 1997).

Some text coverage researchers dismiss the importance of keywords on reading comprehension. Although N. Schmitt et al. (2011) did not use the pseudoword method to determine text coverage, they nevertheless broach the topic of whether all words should be equally weighted when calculating the effects of text coverage. Without reference to any supporting studies, N. Schmitt et al. assert that key words would be easily guessed due to multiple occurrences in the particular text and thus do not warrant special consideration. This is a startling assertion to make for two reasons. The first is that it suggests that unknown words which are easily guessable should not be treated as unknown words. If that is true, then determining text coverage suddenly becomes an even more intractable problem because it is impossible to determine which unknown words fit this criterion. The second reason is that the suggestion of frequency of occurrence dictating the ability to guess the meaning of unknown words from the context in which they appear flies in the face of a large body of research, some of which is intimately connected to text coverage studies, which show that guessing an unknown word’s meaning from context is difficult (Kurnia, 2003) with even ten exposures resulting in successful guessing less than two-thirds of the time (Pellicer-Sánchez & Schmitt, 2010), that guessing correctly is highly dependent on a wide variety of reader factors and text factors (Elgort & Warren, 2014), and that the form-meaning aspect of word knowledge is the slowest to develop during incidental vocabulary acquisition (Hu, 2013).

Keywords do carry more meaning in a text and tend to be relatively infrequent in a general corpus which makes them not only more important in reading comprehension,
but also more likely to be replaced by pseudowords when creating a text with a specific text coverage using this technique. An unbiased examination of the relationship between text coverage and reading comprehension which utilises pseudowords would have to account for the effect of unknown keywords as well. Unfortunately, no existing published study seems to have done that yet.

5.4 Evolution of the text coverage model

The historical roots of the text coverage concept reach back more than a century but the modern incarnation in relation to predicting reading comprehension, at least within L2 English reading as studied by some applied linguists, can largely be traced to Laufer and Sim (1985) who attempted to clarify and quantify two closely related thresholds dealing with the transfer of reading skills from L1 to L2: Cummins’s (1976) Threshold Hypothesis and Clarke’s (1980) Short Circuit Hypothesis. Characterising these hypothesised thresholds was an ambitious and perhaps quixotic task as the proponent of the Threshold Hypothesis specifically described it as not amenable to rigorous definition: “the threshold cannot be defined in absolute terms; rather, it is likely to vary both with the amount of time that is spent through L2 and with the type of cognitive operations that must be expressed through L2.” (Cummins, 1976, p. 24) In contrast, the term threshold does not appear at all in Clarke (1980). Rather, the Short Circuit Hypothesis simply suggests that the source of L2 reading difficulties may be a language problem, but may also be a reading problem:

It is possible that two students with comparable L2 proficiency could produce similar L2 reading behaviors for different reasons: one, because s/he is a poor reader and the other because s/he has not been able to transfer his/her L1 reading skills to the target language. (p. 207)

Although the proponents of the Threshold Hypothesis and the Short Circuit Hypothesis never proposed a quantifiable threshold, Laufer and Sim (1985) undertook a study to characterise a threshold “below which it would be unreasonable to expect foreign readers to apply reading strategies” (p. 406). The process for doing so lacks the rigour needed to be reproducible, even for an exploratory study. For example, in establishing the nature of the putative threshold, participants were required to read academic texts and demonstrate the ability to distinguish between explicit and implicit propositions, recognise author intent, skim for key words, and guess from context. No objective criteria were specified for the participants to be deemed at the threshold other than agreement
between two interviewers that “these students could cope with an academic text of the
nature they were tested on (not perfectly, but well enough).” (p. 408) The number of partic-
pants which comprised this group was only seven. Thus, the sought-after threshold
was operationalised based on the performance of a handful of participants on a subject-
ive measure of reading performance.

After measuring the reading comprehension and reading strategy abilities of a larger
group of participants, a table (p. 408) presents data which demonstrate that mean scores
for three groups, under-, at-, and above-threshold, rise monotonically on both measures.
No statistical tests are reported, nor measures of variance, nor the size of the groups. Fi-
nally, to determine the factors which might explain the data, interviewers observed how
some participants dealt with various reading difficulties and drew conclusions based on
the interviewers’ subjective impressions. The process is not described in sufficient detail
to replicate the conclusions.

The study concluded by asserting that vocabulary knowledge was the primary factor
in determining reading comprehension. Laufer and Sim (1985) was an exploratory study
so the standards of research are necessarily lower, but the lack of precision in conducting
and reporting are troublesome because of the influence this single, relatively small-scale
study has had on following text coverage studies. It is not highly cited, but a follow up
study, Laufer (1989), has been cited roughly four times as often and ushered in the class
of studies to which Hu and Nation (2000) belong: studies which attempt to quantify
the specific level of vocabulary knowledge which characterises the very threshold pur-
portedly found in Laufer and Sim (1985). These studies (e.g. Laufer, 1989; Hu & Nation,
2000; Laufer & Ravenhorst-Kalovski, 2010; N. Schmitt et al., 2011)

all rest on the assumptions laid down in Laufer and Sim (1985): a reading comprehen-
sion threshold exists, the threshold is primarily characterised by vocabulary knowledge,
and that the impact of the latter on the former can be accurately measured and described
in terms of text coverage.

This criticism of Laufer and Sim (1985) should not be interpreted as unique to a
single study. Similar criticism can be directed towards Clarke (1980) which constituted
part of the theoretical framework in and was published just a few years prior. Clarke
sought to determine whether L1 reading skills could be transferred to L2 reading. To
address this question with compelling evidence, one would need to determine whether
L1 reading ability accounts for a measurable degree of L2 reading performance variance
while controlling for L2 ability. Multiple linear regression would be appropriate, but such
an analysis was simply not routine at that time. However, statistical tests of differences
in means were available yet were not employed.
The lasting influence of Laufer and Sim (1985) is surprising when the methodology is compared to other exploratory threshold studies such as Lee and Schallert (1997). In that study L1 Korean reading ability, L2 English reading ability, and L2 English proficiency were compared in 800 Korean middle school and high school participants. The study found that the correlation between L1 and L2 reading ability increased as L2 proficiency increased and buttressed the conclusions with extensive statistical analyses including the determination that L2 proficiency explained almost twice as much variance in L2 reading ability as L1 reading ability did. Lee and Schallert (1997) literally operationalised the threshold hypothesis such that its effect would be detectable through statistical analysis. Yet despite citing that same study, Laufer and Ravenhorst-Kalovski (2010), a recent text coverage study, did not measure L1 ability.

Ironically, Hu (1999) does provide a minuscule ray of hope for the text coverage model, but it is not reported in Hu and Nation (2000). Like in Lee and Schallert (1997), a significant correlation between MCQ and vocabulary size was reported for the 100% text coverage condition, but not for the lower text coverage conditions. This would possibly be indicative of a threshold, however, no significant correlation were found between the cued recall score and any of the text coverage conditions. Given that the variance in the MCQ scores for the 100% text coverage condition was not equivalent to the other conditions, there is little credence that can be given to these data.

This is why it would help to improve existing measures of L2 proficiency before moving further down this path. Text coverage clearly is not the way to measure a threshold in the Cummins and Clarke sense, but by investigating how the relationship between L1 reading proficiency and L2 reading comprehension varies with L2 proficiency, evidence of a threshold may indeed appear to be compelling. But there are few good L2 proficiency measures which are inexpensive, efficient, and reasonably accurate. Putting the VST online would accomplish those goals.

As with the lack of precision in summarising previous research, the concept of a threshold has shifted away from Cummins’s (1976) Threshold Hypothesis. The mechanisms involved in the Threshold Hypothesis and Clarke’s (1980) Short Circuit Hypothesis have also been lost. This can be seen most clearly in N. Schmitt et al. (2011) when graphs illustrating the hypothetical relationship between text coverage and reading comprehension are presented. The are essentially the same in appearance as in Lee and Schallert (1997), but the explanations differ. In Lee and Schallert’s, the y-axis represents the strength of the correlation between L1 reading ability and L2 reading ability, while the x-axis depicts L2 proficiency. This is an accurate representation of what we would expect to see from the Threshold Hypothesis and Short Circuit Hypothesis. In
N. Schmitt et al.’s, however, the $y$-axis is L2 reading comprehension and the $x$-axis is text coverage. In other words, the L1 skills of the readers are not present in the model. In fact, it can not be part of the model because text coverage deals with nothing but the relative L2 proficiency, as measured by vocabulary size, in relation to a given text. Vary the text, and the text coverage experienced by the reader will change, but the reader’s L2 proficiency has not changed. The type of threshold which originally motivated text coverage studies has not been crossed. The determinant of text coverage is the text itself in relation to the reader, not the absolute L2 proficiency level of the reader. Therefore, no change in the ability to draw on L1 reading ability would be expected. When a text is very easy, i.e. the number of word types are low and their frequencies generally high, a reader with low L2 proficiency might experience a high text coverage. Yet the reader’s L1 reading skills would still not be available because of the short circuit caused by low L2 proficiency.

N. Schmitt et al. (2011) conclude by determining that there is no threshold. It would have been impossible to detect one anyway because the variables required to make that determination were not measured. The concept of a threshold in this line of research has morphed to a strictly surface feature whereby L2 (English) reading comprehension should undergo a rapid decline with respect to text coverage after a certain point. The relationship appears more or less linear, not indicative of a threshold yet the conclusion does not describe a linear relationship in terms of units of comprehension increasing by units of text coverage. The conclusion is around 95% text coverage is necessary for a score of 60% on a reading comprehension measure, 98%–99% text coverage required for 70% reading comprehension, and 100% text coverage required for 75% reading comprehension. These are descriptions of thresholds; to have a certain level of performance, a minimum level of ability is required. The description of a linear relationship is later reported as “an increase in comprehension growth of 2.3% for each 1% growth in vocabulary knowledge” (N. Schmitt et al., 2011, p. 39) yet no confidence intervals are given. Acknowledgement is given to the large variance, yet that does not survive in the Implications Section where 98% is recommended to achieve reading comprehension scores $>60\%$. But Figure 4, which features not confidence intervals but standard deviations shows that it’s entirely normal for readers at 90% coverage to achieve $>60\%$ reading comprehension and only at 100% does scoring 60% or less become rare, if normal is defined as $\pm 1$ standard deviation and rare is beyond that.
5.5 Conclusion

Studies which support the text coverage model of reading comprehension typically temper their findings by noting that vocabulary knowledge is not the only factor involved in reading comprehension, but often times this caveat seem perfunctory at best. The following quotation from a footnote in N. Schmitt and Schmitt (2014) regarding the widely cited 98% text coverage figure exemplifies the pattern often seen. It begins with an initial hedge regarding the role of text coverage in relation to other reading comprehension factors, but the weight of the hedge is almost completely offset by the unflagging insistence which inevitably follows that high text coverage will virtually guarantee a significant level of comprehension:

The current consensus is that 98% lexical coverage [...] is necessary for adequate comprehension. Of course, knowing this amount of vocabulary does not guarantee reading comprehension, as reading involves more than just vocabulary knowledge, but research indicates that if readers know enough words to cover 95–98% of a text, they are likely to obtain 60–68% comprehension of that text (pp. 6–7).

In other words, there is no guarantee that text coverage can predict reading comprehension, but usually it does.

Despite the appearance of acknowledging that reading comprehension arises from the confluence of a variety of variables, the conclusions of these studies inevitably make concrete recommendations based on text coverage alone. There appears to be little to no effort to investigate how, nor to what degree, text coverage interacts with the other factors. The impression of confidence in the ability of text coverage to predict reading comprehension is at odds with the paucity of evidence to support it. Those that provide some support, as discussed in this chapter, are often littered with suspicious assumptions and questionable conclusions.

The question that must be answered in this critical review of the text coverage model of reading comprehension is thus: is there enough good evidence to give credence to it? When one steps back and takes an honest, unbiased, unassuming look at the existing evidence and data, a compelling, robust relationship between text coverage and reading comprehension does not stand out. The text coverage model does not seem to be sufficient in theory, nor in supporting empirical evidence, to justify the claims that have been made. The claims themselves may indeed be largely correct, such as I. S. P. Nation’s (2006) assertion that a large vocabulary is necessary for reading with ease, but the actual figures and the reasons for them are likely misleading. The probable truth
5.5. CONCLUSION

is that reading requires a level of language proficiency, of which vocabulary size may be a sufficient proxy, to reach a critical, but broad and fuzzy, threshold of competence such that low level decoding and lexical access become automated enough for sufficient mental resources to be reserved for higher level tasks such as syntactical parsing and construction of mental representations of explicitly encoded textual propositions. It is also likely probable that larger vocabulary sizes also help improve the odds that a reader will know not only a great proportion of the running words in a text, and hence a high text coverage, but also have reasonable knowledge of a high proportion of the key words in that text which carry a greater share of the meaning encoded in the text. Likewise, Cummins’s (1976) Threshold Hypothesis and Clarke’s (1980) Short Circuit Hypotheses both might contain kernels of truth (Lee & Schallert, 1997) but they are simply useful fictions which help shape the conceptualisation of L2 reading and the dynamic nature that various L1, L2, and language-independent factors play in the process. Researchers must be careful to not take that useful fiction literally no matter how convenient or appealing it appears.

All of the text coverage research since Laufer and Sim (1985) has been predicated on the assumption that a quantifiable text coverage threshold exists and blocks comprehension directly. Given the current state of affairs, it would be apropos to move on from the text coverage model of reading comprehension from future research, not because it does not exist, but because it is not useful (Bernhardt & Kamil, 2006). When predicting reading comprehension, existing L2 readability formulae, which usually account for lexical and syntactic complexity, explain much more variance than any of the more methodologically rigorous text coverage studies. Brown (1998), for example, combined four text characteristics to explain 51% of variability on cloze tests, often used as a proxy for reading comprehension tests. The texts were analysed for syllables per sentence, word frequency, proportion of long words, and proportion of function words. A year later, Greenfield’s (1999) Miyazaki EFL index was found to explain 74% of variability on a cloze test using only two characteristics. This index calculates reading difficulty from letters per word and words per sentence and was calibrated against L1 Japanese readers. Even more variability can be explained with Crossley, Greenfield, and McNamara’s (2008) formula: 86% of variability on a cloze test by adding a meaning construction index to measures of lexical and syntactic complexity. No study of the text coverage of reading comprehension has ever explained that amount of variance.

Furthermore, the concept of a text coverage threshold, at least those which approach the threshold concept at a superficial level without regard to L1 reading abilities, should similarly be ignored in future research. In fields such as psycholinguistics, there are...
obvious thresholds where some large shift in effect begins or ceases to occur around
a very narrow range of values. Reaction times, firing neurons, speech rates, and sim-
ilar phenomena all exhibit a classic threshold which is quantifiable and robust. There
are physiological limits in these cases which are well characterised and useful to know.
They can be used for real-world tasks and are valid examples of thresholds. There are
likely a range of situations where deficits in critical factors can severely impede both
the processes and products of reading comprehension, but it is highly unlikely that they
can be characterised as true thresholds in any meaningful way. The field of L2 read-
ing research will be greatly advanced by focusing efforts instead on understanding the
factors involved in reading, the interactions between them, and the effects on various
processes and products of reading comprehension.

This chapter’s highly critical review of the research within the text coverage tradition
exposed serious deficiencies in both the model and the evidence supporting it. It has
attempted to demonstrate that the text coverage concept suffers from an unfortunate
conflation of incompatible applications and methodologies. It is an overly simplistic
view of reading which does not help the field develop a clearer understanding of L2
reading (Bernhardt & Kamil, 2006). As a predictive model of reading comprehension,
it is not skillful enough to justify further study when compared to existing readability
formulae which enjoy a much larger foundation of empirical support.
Chapter 6

Measuring vocabulary size

The broad goal of applied research based on the text coverage model, and specifically on the ideal text coverage figure reported in Hu and Nation (2000), is to match readers to texts in much the same manner as multivariate readability formulae do. This approach attempts to estimate the vocabulary sizes needed to comprehend various types of text (e.g. I. S. P. Nation, 2006; Webb & Macalister, 2013). One assumption these studies make is that texts of a particular genre are that the lexical profiles of texts within the genre are more homogeneous than texts between genres. By grouping texts of a specific genre together in a corpus and averaging across all texts, these studies then calculate and report the vocabulary size a reader would need to have to experience 98%, or 95%, text coverage.

In addition to the methodological objections of the text coverage model raised in chapter 5 (page 111), this approach to applying the model is unfortunately the wrong direction. Given the available technology, and the trajectory of its development, characterising any particular text is a trivial task. Every text included in a corpus created for these studies is amenable to individual analysis; averaging characteristics of texts in a corpus simply obliterates their individual differences. If the goal is to match a specific reader to a specific text, the most important information in achieving that goal will be located in that specific text and that specific reader, not in a corpus. There is no compelling reason to determine the vocabulary demands of a particular genre because that reveals nothing about any particular text within that genre.

Whether or not the text coverage model will continue to garner interest among researchers and practitioners remains to be seen. If it does not, then a strong case can be made for increased focus on measuring the L2 proficiency of readers. Likewise, even if its popularity continues, it is still important to improve our ability to measure L2 proficiency accurately. How could both of these scenarios lead to the same conclusion?
The text coverage model focuses on the overlap between the lexical profile of a text and the mental lexicon of the reader. The former can already be measured with great precision through a variety of free software packages (e.g. Anthony, 2014) and websites (e.g. Cobb, 2009). On the other hand, even multivariate reading models, both explanatory models such as Bernhardt’s (2005) compensatory model of L2 reading and predictive models such as most readability formulae (e.g. MetaMetrics, Inc., 2007), also require precise measures of a variety of reader abilities. One such useful measure could be vocabulary size, i.e. the size of the reader’s mental lexicon. If that is a reasonably proxy of an ability useful in reading, then in both cases a more robust and precise measure of vocabulary size would be invaluable as it would help to more accurately characterise a particular reader when matching the reader to a particular text.

Given the results of the replication and expansion of Hu and Nation (2000) described in chapter 3 (page 59) and the concerns for the viability of the text coverage model discussed in chapter 4 (page 97), it seems that the best course of further research in this thesis is to focus on more accurately characterising the most imprecise part of the equation: the reader. For the field of applied linguistics as a whole, a standard measure of vocabulary size which is cheap to administer, both in terms of time and cost, would increase the likelihood that participant vocabulary size would be reported as a matter of course. This would have the knock-on effect of making these studies more amenable to meta-analysis by making the studies’ participants more directly comparable.

Finally, a sharp shift in direction is warranted here because scientific inquiry should follow the data as they present themselves. Researchers should be completely agnostic and willing to abandon preconceived notions when the data suggest errors in thought or assumptions. If a different factor holds the potential of better predicting reading comprehension, which is the ultimate goal of this research, then it is incumbent upon the research to investigate it further. In this case, the VST seemed to predict a substantial amount of reading comprehension variance. If the VST can be improved in any way, such as in reducing error or increasing precision, then it logically follows that the predictive power of that measurement may well increase as well.

There is already an abundance of evidence supporting the predictive power of vocabulary size in reading comprehension, so vocabulary size tests are obviously valuable and important to understanding reading performance. Despite that fact, vocabulary size test development is still immature (P. Nation, 1993) which is manifest in large L2 English reading studies misusing outdated tests. Laufer and Ravenhorst-Kalovski (2010), for example, used the VLT, which was created as a diagnostic test (P. Nation, 1983), to estimate vocabulary size despite the availability of the VST which was specifically cre-
ated to measure that construct (P. Nation & Beglar, 2007). Clearly there is a need for an accurate, efficient, and standard measure of vocabulary size along with age and L1 norms to use as baselines for comparison. Further developing a test such as the VST and making it available to researchers and practitioners alike would likely have a much greater impact on L2 English reading comprehension research, and hopefully other applied linguistics research, than further investigations into the moribund text coverage model of reading comprehension.

6.1 Challenges in measuring vocabulary size

Vocabulary size has been measured for over 100 years but very little methodological progress was made for most of that time period (P. Nation, 1993). Thorndike (1924), for example, assessed the measurement methodology of vocabulary tests at that time and detailed serious issues that needed to be avoided. Unfortunately the advice on how to improve the situation was largely ignored for the next few decades. The fallacious assumptions and methodologies which impeded advancement of valid measures of vocabulary size have still not been eradicated; they persist to this day.

In 2011, for example, the “The One Show” interviewed David Crystal, a highly regarded linguist and prolific author who explained that one’s vocabulary size can be estimated quite simply and expeditiously by selecting a few random pages from the dictionary, reading through the entries and tallying the number of known words and entries, then multiplying the total number of entries as reported by the publisher by the ratio of known words to entries. For the younger guests on the show, who used typical student-sized dictionaries to apply this method, the range of vocabulary sizes they derived was in the range of 10,000 to 30,000 words. The host of the segment was then shown calculating his own vocabulary size, this time using a large, historical dictionary and discovered his vocabulary size to be greater than 100,000 words.

What the producers of “The One Show” failed to realise is that the segment had clearly demonstrated a phenomenon that had been known for more than a century commonly known as the **big dictionary effect**. A full discussion of the pitfalls potential vocabulary size test creators should avoid can be found in P. Nation (1993), but the moral of the story is that accurately estimating vocabulary size is not a simple task and going on intuition and logic alone is insufficient. The **big dictionary effect** seems to be one of the most pervasive pitfalls in measuring vocabulary size but to avoid it, a reliable frequency list needs to be used in place of a dictionary.
The VLT (P. Nation, 1983; N. Schmitt et al., 2001) is another good example of how good intentions in measuring vocabulary size have gone wrong. The VLT was not called a size test for a reason. It was designed as a diagnostic test to aid teachers and learners in focusing their vocabulary learning efforts on the most frequent words which were still unknown. If the test showed full mastery of the 2000 and 3000 levels, but not the 5000 or 10,000 levels, then based on the idea that the most frequent words are also the most useful, a learner’s time could be best utilised by focusing on words in the 4000 to 7000 range before spending too much time on the more difficult, and less useful words near the 10,000 range (P. Nation, 1983). The problem is that researchers have taken the VLT and often reinterpreted it as a vocabulary size test (N. Schmitt, 2014). At first blush, there does not seem to be any reason why this would not be a good idea, but as “The One Show” example demonstrates, intuition in testing does not always lead to valid results.

6.2 Limitations of the VST

The VST (P. Nation & Beglar, 2007) represents a large step in the right direction. Groups of 1000 word families with similar BNC frequencies created by I. S. P. Nation (2006) are used as the basis for the VST items. There are 10 VST items for every 1000 word family grouping, so each correct response implies knowledge of 100 word families of similar frequency. Each MCQ items is presented with a short sentence that provides very little semantic context. The example sentence and the four options are comprised of only high frequency words to prevent the test itself becoming a barrier to accurate measurement. Despite all of these advances over other existing tests, there are still many limitations and shortcomings which show potential for correction.

6.2.1 Item ordering

The test items are ordered according to frequency; items representing the most frequent word families appear first and the least frequent word families, consequentially the most difficult, appear at the end of the test. Although this ordering may be beneficial for some purposes, it may lead to underestimating the participants’ true vocabulary size. As the test can take up to twenty minutes to complete, fatigue can set in and the degree of effort put forth on the part of the participants could decline at the time when it is most needed.

The response patterns of some participants in the approximate external replication study (see subsection 3.3.6, page 82) suggest that some participants did indeed attenuate
the effort put into responding to VST items after a certain point and simply selected options at random. Observation by the researcher during administration of the VST indicated that this cessation of effort may have been attributable to the step-wise increase in difficulty inherent in this ordering scheme.

It is incontrovertible that less frequent lexical items are less likely to be known, but the degree to which this is true on the VST may a function of the item order. If participants expend less effort as a result of previous items, local independence would sacrificed. This would effectively increase the difficulty of items appearing later in the test solely as a consequence of their position in the test rather than the inherent difficulty of the item being tested. Uncertainty about accuracy of the vocabulary size estimate would increase if such a phenomenon occurred. When items are ordered in descending frequency, the issues of fatigue may confound the difficulty of lower frequency items and confound the results. The result would be manifest in a underestimate of vocabulary size.

It is imperative to make the vocabulary size estimate as accurate as possible. If a reader’s vocabulary size is used to estimate text coverage for a particular text, for example, then underestimation would result in a lower text coverage figure. Randomisation should help restore some local independence, but as the VST was designed as a paper-based test, this is difficult to implement in its present format. Multiple versions could conceivably be created with random item orders, but it is impractical to do so for each participant. To remove any item ordering effects, every item in every test used to establish norms should be randomised.

6.2.2 Completion time

Another observation made by the researcher during the approximate external replication study (chapter 3, page 59) regarding the VST was that some participants completed the VST relatively quickly whilst others lingered on each question, revised previous answers, marked others for further consideration later, and/or used an inordinate amount of time to complete the test. Yet it is possible, and perhaps even common, for two participants to receive exactly the same estimates of vocabulary size while taking radically different lengths of time to complete it. The paper-based VST takes no account of these differences.

1Given the severity of methodological issues raised in chapter 4 (page 97) concerning the text coverage model of reading comprehension, this procedure is not recommended.
This type of problem was noted almost 100 years ago when Cuff (1930) criticised existing vocabulary tests as “planned to find out how many or how difficult words one can define, rather than how accurately or how rapidly one can define them.” (p. 219)

The importance of the speed of defining vocabulary comports with C. Perfetti’s (2007) Lexical Quality Hypothesis which posits that the quality of lexical item storage, i.e. the speed at which a lexical item can be recalled, the grammatical classes and inflections are known, and the ability to disambiguate it from similar items, all have significant effects on reading comprehension. C. A. Perfetti and Hogaboam (1975), for example, found that naming speeds differed in poor and good comprehenders, even for pseudowords. This implies that the speed of access, or decoding speed for pseudowords, is important to reading comprehension, not just the definition.

6.2.3 Test-taking strategies

As with any MCQ test, it is possible to guess, either blindly or through the process of elimination, on unknown items in an attempt to increase one’s overall score. Various schemes can be employed to detect and correct for this behaviour such as penalising incorrect responses using the three parameter model of item response theory, which sets the lower asymptote of the item characteristic curve as close to the inverse of the number of options. There is some controversy over which correction methodology is most appropriate as incorrect application can result in unjust penalties. Ben-Shakhar and Sinai (1991), for example, point out that guessing strategies are not only contingent upon situational factors such as time constraints and test instructions, but also on subject factors such as gender. Although the difference is small, the findings suggest that females are more likely to not answer some items rather than guess even when given permissive test instructions which encourage guessing and the test is marked using scoring which does not penalise incorrect answers (number right scoring).

Even when formula scoring or other types of correction are employed, they are only applicable at a superficial level. They do not take into account the order in which the items were answered on a paper-based test where order is not enforced nor the amount of time spent employing a strategy on a particular item. Strategy usage may not be consistent throughout the test or across items. More time might be invested in the middle of the VST where the difficulty of the items closely matches the limits of the participant’s vocabulary knowledge, possibly indicating a great deal of pondering on the part of the participant. An inordinately short amount of time might then be spent near the end of a test either because the allotted time is almost expended or the participant
has given up. Although adjustments made to the final size determination may lead to a more correct estimate than raw scores, it is possible that applying corrections at the item level, as a function of time spent answering a particular item to indicate whether the participant was randomly guessing or not, may produce even more accurate estimates.

### 6.2.4 Over- and under-testing

The VST was designed to accurately estimate vocabulary sizes of a wide range of ability levels. The trade-off of this design goal is that lower level participants will be forced to answer dozens of items which far exceed their ability if the entire test is given; higher level participants will similarly face the prospect of responding to a significant number of items which are far below their ability level and may appear to be a waste of time or even condescending. For those with expansive vocabularies, this protocol could become both tedious and tiring especially when items are presented in order of frequency.

Shortening the test by removing the lowest or highest frequency words is an option to address this issue, but care must be taken to not over- or underestimate the participants’ ability. Without actually measuring that ability in the first place, the probability is high that the test administrator will commit that very error. It is probable that some participants in the approximate external replication study (chapter 3, page 59) had larger vocabulary sizes than estimated by the VST because its length was shortened to just the first eight levels. Ideally participants should be tested on word families at increasingly infrequent levels until she is unable to answer at a rate higher than chance. Beglar (2010) recommends having participants sit at least two levels beyond their ability level, but as noted in the previous section, it is difficult to know in advance what that level might be.

A computer-adaptive test would help ameliorate this problem by continuously adjusting the difficulty of test items until a confident estimate is obtained. This can only be implemented after a significant amount of data has been collected for each item to determine item difficulty and related characteristics. In order to facilitate the data collection, a simple computerised non-adaptive test would be useful as the first step towards an adaptive test.

### 6.3 Research Questions

To improve on the current limitations of the VST, the following research questions will be addressed:
The accuracy of the VST is contingent upon participants maintaining focus and effort for more than 100 items which increase monotonically in difficulty every 10\textsuperscript{th} item. It is hypothesised here that presenting at least some of the more difficult items near the beginning of the test and reserving some of the easier items for the end of the test may result in more accurate vocabulary size estimates. Therefore, Research Question One is:

**Research Question One:** Does the order of item presentation in the VST have an effect on test performance?

The wide variety of completion times should be a function of some vocabulary-related ability which is not captured by the paper-based VST. Since speed of lexical access is likely to have an effect on tasks such as reading comprehension, Research Question Two is:

**Research Question Two:** Does any time component account for any variance performance on VST?

Standardised tests used in psychology and education are useful not only because they have been piloted and revised, but also because norms of performance have been established with which to compare new results. This is something which is sorely lacking in the measurement of vocabulary acquisition. There simply are not many studies which have reported vocabulary size estimates of large populations and those that do exist often differ in the tests used, conditions under which the tests were administered, and other critical factors. Norms for each L\textsubscript{1} group, delineated by age and, ideally, other important demographic characteristics as well, would address this deficiency. For example, when norms are established, teachers could use the VST to locate their students relative to other learners and identify those who may need additional assistance in improving vocabulary breadth. Likewise, learners who are found to have vocabulary sizes well above the norm may better allocate their time by focusing on fluency or production rather than breadth. Therefore, Research Question Three is:

**Research Question Three:** What are the performance norms of the VST for various L\textsubscript{1} learner groups of various ages?

### 6.4 Refining the VST

In order to facilitate the refinements detailed below, the format of the VST was converted from a paper-based test to an online, computerised test located at http://my.
vocabularysize.com. Initial work on the conversion was done in collaboration with a group of undergraduate software engineering students at Victoria University of Wellington. This stage of development consisted of bi-weekly group programming sessions where the researcher met with the students to discuss requirements, make implementation decisions, and learn how the software framework was designed.

The code that runs on the HTTP server is written in PHP on top of the Jade Framework (Hindmarsh, Calcinai, & Renner, 2010). The website is backed by a MySQL database where the information related to the tests, such as the test items, results, and various settings, are stored. HTML forms capture participant responses and a web browser submits the data to the HTTP server where the data are stored in the database for later analysis. At the completion of a test, general demographic information is collected from the participant including age, gender, $l_1$, time spent learning the target $l_2$, time spent living in an $l_2$-speaking country, and general comments. No questions are mandatory; “no answer” is an available option for each question. The VST results are presented to the participant after submitting answers to these questions.

Another justification for making the VST available online is to make the test more useful to teachers and researchers. As a paper-based test with more than 100 items, marking by hand can become prohibitively expensive in terms of time as the number of participants increases. A facility for teachers was incorporated into the website to make group administration of the test possible. This facility provides teachers with the ability to create a customised test session for a class or group. By doing so, questions which require PII, such as participants’ names or identification numbers, can be added to the beginning or conclusion of the test. This additional information can then be used to identify to whom a particular result belongs. The results are stored separately from PII in the database so that teachers and researchers can both access item response data while preventing unauthorised access to PII associated with a specific data set. Only the creator of a customised test session can view responses to questions added to that session and connect them to the results.

The course, SWEN 302 — Agile Methods, required students to work together on a software project to learn the process of agile programming by experiencing the project firsthand. Although they made extraordinary progress in creating the initial adaptation of the VST from paper to website, completion of the project was not a course requirement. All of the refinements discussed in this chapter were implemented by the researcher as modifications and additions to the initial work carried out with the SWEN 302 students.

The Jade Framework was not publicly available when the VST conversion done. Its name was changed to Wave Framework when it was placed on GitHub (http://github.com).
A positive side effect of group administration is that the results are less likely to suffer from bias inherent in data collected from self-selected populations. The anonymised results contain a numeric key by which results can be grouped if the data were collected as part of a group administration. These subsets represent less biased and more generalisable data which can then be used as a baseline with which to compare the larger public data set for bias.

### 6.4.1 Item presentation order

Initially the website was designed to simply present each item in the same order as the paper-based version to get a baseline reading on performance. To vary the presentation order such that items representing every difficulty level would be presented to the participant at every stage of the test administration, the VST items were divided into seven groups, each containing twenty questions drawn from two adjacent frequency levels. This grouping was selected because the only published validation study of VST determined that it can only discriminate between seven distinct levels (Beglar, 2010). The questions in each group were then rearranged such that every 20 consecutive items, i.e. seven consecutive stages, comprised items from each of the seven consolidated groups. The ratios of items in each stage varied such that the average frequency level of all the items in each stage increased compared with the preceding stage.

The effect of this ordering ensured that participants have an opportunity to answer at least one question from each of the consolidated groups within the first twenty questions. Similarly, in the last twenty questions, the participant will encounter at least one question from the first consolidated group. The proportions of questions from each of the consolidated groups which appear in each of the stages is summarised in the Table 6.1 (page 157). The effect of this algorithm is displayed graphically in Figure 6.1 (page 157) where the probability of each of the original VST items appearing at a particular sequence in the test administration is plotted as a contour map. The items are accordingly randomised each time a new VST session is begun which ensures that item ordering effects are eliminated. For language practitioners, a practical benefit of this algorithm is that the potential for copying responses from neighbouring learners is severely diminished.

Although the items were randomised within each of the stages, the first item of each stage was constrained to match the original consolidated group, i.e. the first question of the third stage was always drawn from the third consolidated group. This ensured that

4Examples of this bias can be found in section 6.7 (page 168).
Table 6.1: Proportions of original VST items in each of the new groupings for the online VST.

<table>
<thead>
<tr>
<th>Online stage</th>
<th>Original item frequency levels</th>
<th>1–2</th>
<th>3–4</th>
<th>5–6</th>
<th>7–8</th>
<th>9–10</th>
<th>11–12</th>
<th>13–14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–20</td>
<td>30% 20% 15% 10% 10% 10% 5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21–40</td>
<td>20% 20% 20% 10% 10% 10% 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41–60</td>
<td>15% 15% 20% 20% 10% 10% 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61–80</td>
<td>10% 15% 15% 20% 15% 15% 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81–100</td>
<td>10% 10% 10% 20% 20% 15% 15%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101–120</td>
<td>10% 10% 10% 10% 20% 20% 20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>121–140</td>
<td>5% 10% 10% 10% 15% 20% 30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.1: Original VST item presentation order probability density in online version.
the very first question of the test was always a very high frequency item and would not frustrate the participant from the outset.

This pseudo-random mixing algorithm yielded an ordering which, on the whole, increased in difficulty as the test progressed so that the participant would sense that the items were becoming more difficult, while reserving some high frequency items for the final stage to prevent the participant from giving up in total frustration, thus addressing the limitation of item ordering.

6.4.2 Decision time

Almost every model of reading either assumes fluent word recognition or automaticity of sight vocabulary. In fact, some theories of reading difficulties suggest that one of the primary causes may be that the reading process is impeded by slow word recognition. This forces readers to expend too much time and working memory capacity on those lower-order processes and consequentially leaves insufficient resources remaining to perform higher-order comprehension processes. One example is C. A. Perfetti and Hogaboam’s (1975) verbal efficiency hypothesis which posits that many of the reading subtasks must be automated because they have to share such a limited working memory capacity, the bulk of which needs to be allocated to higher-order reading tasks.

Traditionally, word recognition time is measured using a lexical decision task where the participant looks at a monitor displaying a string of letters. The participant then indicates whether the string appears to be a known word or not by pressing a button, indicating the decision as quickly as possible. The time it takes to make this decision is called the reaction time. Shorter times indicate more mature and rapid lexical access. Common words are recognised faster than uncommon words because it takes some time to decide whether a less frequent word is simply uncommon yet known, or completely unknown. Pseudowords, strings of letters which are orthographically and phonologically valid, are also presented as controls.

The timings recorded in the revised VST do not provide the level of sensitivity found in lexical decision task reaction times because participants must read more than a single word. In the best case scenario, a participant would read the target word and at least one option before making a decision, assuming the first option was the correct answer. On average, however, participants must read at least two options and spend time positioning the cursor over the correct option before clicking. It is also likely that the participants read some or all of the example sentence, which is intended to provide some syntactic context without revealing meaning, as well. Because of the large amount of reading
which must be performed before recording a response, the time taken to complete the
task should more properly be referred to as decision or response time rather than reac-
tion time.

The decision time should be a function of the number of words which comprise the
prompt, most of the example sentence, and half of the total words in the options. This
is based on the assumption that participants are likely to read less than half for very
high frequency items, but perhaps every word for low frequency, or difficult, items. On
average, participants are unlikely to read every word of every item, but the actual pro-
portion of words that are actually read is unknown. In a way, measuring decision time
in this manner is not simply measuring the speed of access that reaction time would
measure. It is confounded by reading ability itself, so there should be a moderate correl-
ation between decision time and reading comprehension, especially if participants are
couraged to complete the tasks expeditiously.

The time taken to respond to each item was recorded with millisecond resolution by
JavaScript, a programming language which is embedded in a web page and is executed
within the browser. The timing began when the item was presented and ended when the
participant made a selection. The time was appended to the HTML form before submit-
ting the response to the server. Although some participants completed the test without
JavaScript enabled in their browsers, thereby preventing the timer from operating, the
vast majority of response sets include this measurement.

6.4.3 Immutable decisions

Unlike the paper-based test, the revised version of the VST does not permit participants
the opportunity to revisit or reconsider any previous answers. Although this restric-
tion may be suboptimal for some participants, it essentially forces all of the participants
into similar testing strategies. One strategy this restriction prevents is that of skipping
all items which are not immediately answerable with confidence in order to reconsider
them with any remaining time later. The strategy applied at such a time might include
process of elimination to improve the odds of successfully guessing which would seri-
ously undermine the fidelity of the test. It is possible that participants who do choose
to employ said strategy in the course of the test might be identified by longer decision
times for particular items although such a conclusion would only be inferential.
### Table 6.2: Participant L1s

<table>
<thead>
<tr>
<th>L1</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>31,513</td>
</tr>
<tr>
<td>Russian</td>
<td>20,829</td>
</tr>
<tr>
<td>Chinese (Mandarin)</td>
<td>19,899</td>
</tr>
<tr>
<td>Japanese</td>
<td>9,255</td>
</tr>
<tr>
<td>Arabic</td>
<td>3,998</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>2,564</td>
</tr>
<tr>
<td>Indonesian</td>
<td>2,469</td>
</tr>
<tr>
<td>Spanish</td>
<td>2,057</td>
</tr>
<tr>
<td>German</td>
<td>2,012</td>
</tr>
<tr>
<td>Korean</td>
<td>1,949</td>
</tr>
<tr>
<td>Persian</td>
<td>1,754</td>
</tr>
<tr>
<td>Turkish</td>
<td>1,713</td>
</tr>
<tr>
<td>Ukrainian</td>
<td>1,361</td>
</tr>
<tr>
<td>Dutch</td>
<td>1,341</td>
</tr>
<tr>
<td>Finnish</td>
<td>1,336</td>
</tr>
<tr>
<td>Polish</td>
<td>1,327</td>
</tr>
<tr>
<td>Chinese (Yue)</td>
<td>1,310</td>
</tr>
<tr>
<td>French</td>
<td>1,120</td>
</tr>
<tr>
<td>Hindi</td>
<td>1,117</td>
</tr>
<tr>
<td>Malay</td>
<td>970</td>
</tr>
<tr>
<td>Filipino</td>
<td>940</td>
</tr>
<tr>
<td>Portuguese</td>
<td>929</td>
</tr>
<tr>
<td>Thai</td>
<td>825</td>
</tr>
<tr>
<td>Czech</td>
<td>801</td>
</tr>
<tr>
<td>Italian</td>
<td>650</td>
</tr>
<tr>
<td>Tagalog</td>
<td>523</td>
</tr>
<tr>
<td>Tamil</td>
<td>490</td>
</tr>
<tr>
<td>Urdu</td>
<td>477</td>
</tr>
<tr>
<td>Telugu</td>
<td>403</td>
</tr>
<tr>
<td>Romanian</td>
<td>347</td>
</tr>
<tr>
<td>Chinese (Wu)</td>
<td>339</td>
</tr>
<tr>
<td>Chinese (Xiang)</td>
<td>332</td>
</tr>
<tr>
<td>Bengali</td>
<td>270</td>
</tr>
<tr>
<td>Estonian</td>
<td>256</td>
</tr>
<tr>
<td>Malayalam</td>
<td>230</td>
</tr>
<tr>
<td>Marathi</td>
<td>187</td>
</tr>
<tr>
<td>Chinese (Min Nan)</td>
<td>151</td>
</tr>
<tr>
<td>Chinese (Jin)</td>
<td>119</td>
</tr>
<tr>
<td>Chinese (Hakka)</td>
<td>119</td>
</tr>
<tr>
<td>Punjabi</td>
<td>119</td>
</tr>
<tr>
<td>Chinese (Gan)</td>
<td>118</td>
</tr>
<tr>
<td>Gujarati</td>
<td>109</td>
</tr>
<tr>
<td>Javanese</td>
<td>107</td>
</tr>
</tbody>
</table>
Table 6.3: Location of participants

<table>
<thead>
<tr>
<th>Participant location</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>23,889</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>14,203</td>
</tr>
<tr>
<td>Japan</td>
<td>9615</td>
</tr>
<tr>
<td>China</td>
<td>8675</td>
</tr>
<tr>
<td>New Zealand</td>
<td>7175</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>6301</td>
</tr>
<tr>
<td>India</td>
<td>3874</td>
</tr>
<tr>
<td>Canada</td>
<td>3838</td>
</tr>
<tr>
<td>Ukraine</td>
<td>3720</td>
</tr>
<tr>
<td>Australia</td>
<td>3356</td>
</tr>
<tr>
<td>Indonesia</td>
<td>2549</td>
</tr>
<tr>
<td>Vietnam</td>
<td>2393</td>
</tr>
<tr>
<td>Taiwan</td>
<td>2143</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>2030</td>
</tr>
<tr>
<td>Korea, Republic of</td>
<td>1915</td>
</tr>
<tr>
<td>Germany</td>
<td>1708</td>
</tr>
<tr>
<td>Turkey</td>
<td>1669</td>
</tr>
<tr>
<td>Philippines</td>
<td>1481</td>
</tr>
<tr>
<td>Singapore</td>
<td>1456</td>
</tr>
<tr>
<td>Finland</td>
<td>1426</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1394</td>
</tr>
<tr>
<td>Malaysia</td>
<td>1268</td>
</tr>
<tr>
<td>Poland</td>
<td>1261</td>
</tr>
<tr>
<td>Iran, Islamic Republic of</td>
<td>1222</td>
</tr>
<tr>
<td>Thailand</td>
<td>1081</td>
</tr>
<tr>
<td>Belarus</td>
<td>885</td>
</tr>
<tr>
<td>Qatar</td>
<td>858</td>
</tr>
<tr>
<td>Kazakhstan</td>
<td>822</td>
</tr>
<tr>
<td>Belgium</td>
<td>759</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>723</td>
</tr>
<tr>
<td>Spain</td>
<td>706</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>699</td>
</tr>
<tr>
<td>Brazil</td>
<td>640</td>
</tr>
<tr>
<td>Sweden</td>
<td>609</td>
</tr>
<tr>
<td>France</td>
<td>608</td>
</tr>
<tr>
<td>Hungary</td>
<td>595</td>
</tr>
<tr>
<td>Pakistan</td>
<td>593</td>
</tr>
<tr>
<td>Switzerland</td>
<td>558</td>
</tr>
<tr>
<td>Egypt</td>
<td>520</td>
</tr>
<tr>
<td>South Africa</td>
<td>505</td>
</tr>
<tr>
<td>Italy</td>
<td>485</td>
</tr>
<tr>
<td>Denmark</td>
<td>419</td>
</tr>
<tr>
<td>Norway</td>
<td>409</td>
</tr>
<tr>
<td>Israel</td>
<td>408</td>
</tr>
</tbody>
</table>
6.5 Participants

Participants in this study were largely self-selected. The website is publicly available and there are no restrictions on use. It is impossible, therefore, to verify participant responses to the demographic questions such as their L1 backgrounds. The most popular L1s reported are summarised in Table 6.2 (page 160). One characteristic which can be confirmed with reasonable certainty is the participants’ location at the time the data were collected. This is done by comparing the participants’ IP address to a database of geographic coordinates associated with most available IP addresses. The accuracy of the database at the city or province level is good, but at the country level, it is close to perfect. By comparing participants’ claimed L1 background with their current locations, additional credence can be given to cases where the two factors are in agreement. A list of the countries where participants could be located can be found in Table 6.3 (page 161).

6.6 Results and analysis

6.6.1 Detecting legitimate responses

Several types of datum were collected to aid in the detection of illegitimate datasets:

**test begin and end time:** this enables a check of the total time taken to complete the entire test. Taking an inordinate amount of time, such as two standard deviations from the mean for a particular demographic group, could be due to the participant using supplemental aids to verify definitions or simultaneously performing other tasks and thus not allocating their undivided attention to the test.

**session identifier:** this is a unique random string assigned to each web browser when the VocabularySize.com website is visited for the first time. HTTP is a stateless protocol, *i.e.* there is no inherent way for the HTTP server, which responds to requests from the browser, to uniquely identify a browser and the history of its requests. To get around this limitation through the use of a cookie, a small text file is created by the browser with information from the HTTP server. A string unique to each browser can be set by the HTTP server which will then use that string to associate data stored on the server with the browser request. This makes it possible to personalise responses to the browser requests. For servers capable of delivering content created with PHP, the `session_start()` command automatically creates, assigns, and stores a unique identifying string, called `PHPSESSID`, in a cookie (The PHP Group, 2014).
For each completed test, the first four characters of the participant’s PHPSESSID ID was recorded. Although the actual value of PHPSESSID is 26 characters in length, saving only the first four provided a balance between uniquely identifying user sessions and minimising storage requirements. The string is constructed from a set of 64 possible characters\(^5\) which are selected at random with replacement. Four characters therefore yields \(16,777,216\) possible permutations\(^6\) which was more than enough given the anticipated number of responses.

**individual item timing:** JavaScript, a programming language which runs in web browsers to make web pages interactive, was used to record the time in milliseconds from the initial display of each item until an option was selected. Since JavaScript is enabled on the vast majority of web browsers in the world, almost every data set includes this information.

After examining the data and observing typical behaviour on the VocabularySize.com website through a real-time analytics programme, it appears that there are three broad approaches participants adopt when taking the online VST:

**genuinely sincere:** This group comprises the vast majority of the datasets. The participants appeared to read the instructions at the beginning of the test and, for the most part, follow them to the best of their ability. It is the data generated from this group which is of interest in this study. Only the datasets from this group were included in the analysis.

**rapid clickers:** This group represents about 0.77% of all completed tests. The members of this group appeared to simply click indiscriminately for whatever reason. Their pattern of behaviour shows up in the data set as having abnormally short individual item times as recorded by JavaScript. Their test performance is close to chance, as expected, and the pattern of responses is clearly detectable when plotted against all participants as shown in Figure 6.2 (page 164). It shows that the median response time for items answered correctly lies somewhere between four and five seconds with virtually no participants maintaining that rate below about two seconds per item. Below that threshold, however, another curve begins and peaks around 500 milliseconds which is just shy of double the 260 ms median reaction time in situ-

\(^5\)The possible characters are: abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789,-

\(^6\)The formula for calculating permutation with replacement is \(P^R(n, r) = n^r\) where \(n\) is the population size of elements from which the string is formed (e.g. 64) and \(r\) is the number of positions which must be filled (e.g. 4). Thus, \(64^4 = 16,777,216\).
CHAPTER 6. MEASURING VOCABULARY SIZE

Figure 6.2: Distribution of median decision times for correct item responses.

Table 6.4: Comparison of median correct response decision times for L1 English users

<table>
<thead>
<tr>
<th>Test</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Welch Two Sample t-test</td>
<td>( t(2239.19) = 2.33, p = .020, d = NA )</td>
</tr>
</tbody>
</table>

ations where participants are actively trying to react as quickly as possible (Human Benchmark, 2016). This curve is believed to represent the typical decision times of participants in this group.

repeaters: This group comprises those who take the same test more than once within a short time period and represents about 2.68% of all responses. Their goal appears to be achieving a higher result or learning unknown words from the first round. The participants in this group can usually be identified by repeated occurrences of the same session identifier in the database, but this is not foolproof. There are situations, such as when a school’s computer lab is used by successive groups of students to take the test, where a duplicate session identifier may not indicate a repeater. Their incidence in the data set is higher than the rapid clickers, but still extremely low.
6.6. RESULTS AND ANALYSIS

6.6.2 Item ordering effects

The data indicate that there is a significant ($p = 0.027$) difference in the median correct decision time per correctly answered item of $\approx 300ms$. This means that, ceteris paribus, the stratified pseudo-random presentation format leads to a faster test completion time: 12.2 minutes for the modified presentation order and 12.9 minutes for the original serial presentation order, a difference of $\approx 5\%$. The statistical test results of this comparison can be found in Table 6.4 (page 164). This suggests that harder questions are being answered when the mind is still relatively fresh but that the correctness of the response may not be significantly affected.

6.6.3 The time factor

Despite the crudeness of the measure, median correct decision time was found to account for a statistically significant amount of vocabulary size variance ($\approx 12\%$) for L1 English users. This is a correlation, so no causation is implied, but one could easily understand how the two could be related; participants with large mental lexicons are more likely to have more mature, well-developed mental lexicons. This would result in the faster lexical access demonstrated in this relationship.

Surprisingly, no relationship was found between vocabulary size and median correct decision time for L2 English participants. A comparison of various L1 background participants’ vocabulary sizes regressed on median correct decision time can be found in Table 6.5 (page 166). While unexpected, this lack of a relationship in L2 readers may actually shine light on why the amount of reading comprehension variance explained by text coverage decreased, and vocabulary size explained substantially more, in the replication and expansion study chapter 3 (page 59). It may be possible for L2 English users to do what L1 English users generally can not: acquire a large vocabulary without simultaneously increasing the speed at which the mental lexicon is accessed.

This may be a crucial difference between L2 and L1 reading. For literate, internet-using adults, it may be impossible to not develop automaticity in L1 word recognition simply because of the volume of reading performed on a daily basis. In L2 reading, however, vocabulary size can be increased through intentional learning, a widely promoted technique, without simultaneously improving the speed of lexical access. If this is actually the case, then it would not be surprising that neither text coverage nor vocabulary size explain more than half of the variance in L2 reading comprehension.
Table 6.5: Comparison of the relationship between English vocabulary size and correctly answered item decision times for L1 English users, L1 Russian users, L1 Chinese (Mandarin) users, L1 Japanese users, and L1 Arabic users.

<table>
<thead>
<tr>
<th></th>
<th>L1 English</th>
<th>L1 Russian</th>
<th>L1 Chinese (Mandarin)</th>
<th>L1 Japanese</th>
<th>L1 Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>9239.1857*</td>
<td>6661.6500*</td>
<td>4356.2238*</td>
<td>5373.1452*</td>
<td>10341.9427*</td>
</tr>
<tr>
<td></td>
<td>[9094.0451; 9384.3263]</td>
<td>[6537.5839; 6785.7160]</td>
<td>[4254.5889; 4457.8586]</td>
<td>[5162.6587; 5583.6318]</td>
<td>[9935.1583; 10748.722]</td>
</tr>
<tr>
<td>word.families.total</td>
<td>−0.3334*</td>
<td>−0.0839*</td>
<td>0.0499*</td>
<td>−0.0155</td>
<td>−0.0944*</td>
</tr>
<tr>
<td></td>
<td>[−0.3459; −0.3209]</td>
<td>[−0.0984; −0.0693]</td>
<td>[0.0374; 0.0623]</td>
<td>[−0.0436; 0.0127]</td>
<td>[−0.1632; −0.0255]</td>
</tr>
<tr>
<td>R²</td>
<td>0.1197</td>
<td>0.0064</td>
<td>0.0031</td>
<td>0.0001</td>
<td>0.0030</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.1196</td>
<td>0.0063</td>
<td>0.0031</td>
<td>0.0000</td>
<td>0.0026</td>
</tr>
<tr>
<td>Num. obs.</td>
<td>19977</td>
<td>19810</td>
<td>19780</td>
<td>8229</td>
<td>2382</td>
</tr>
</tbody>
</table>

* o outside the confidence interval
6.6. RESULTS AND ANALYSIS

6.6.4 Learner norms

Research Question Three asked about norms for the VST. The availability of sufficient data to calculate reliable norms for a particular language community relied solely on the popularity of the online test website within that community. At the time of publication, enough data had been collected to calculate a range of age-based norms for just ten languages. The criterion for inclusion in this list is \( \geq 2000 \) complete data sets. The criterion for calculating norms for an age group is \( \geq 20 \) for larger data sets (\( \geq 10,000 \) data sets) and \( \geq 10 \) for smaller data sets.

All age groups represent the age floor of the participants, i.e. a participant who was 10 years 11 months old at the time the test was taken is reported in the 10-year-old age group. It should be noted that any changes in mean vocabulary size between age groups for L2 English participants can not be attributable just to the normal course of vocabulary acquisition. Language policies and learning opportunities within each language community change over time and those changes will be reflected in the norms. This can be seen in some language groups where older participants may have much lower or higher vocabulary size norms than would be expected from an uninterrupted course of vocabulary acquisition. As such, the norms presented here are only valid for a limited amount of time. Ideally data should be collected continuously and the norms updated every few years.

For each L1 group, both a table and plot are provided. The tables specify the mean vocabulary size by age in years, standard deviation, and number of data sets used to make those calculations. To better visualise the data in the table, a plot is also presented which indicates, for each age group, the 50% quantile vocabulary size as a red dot. Grey error bars indicate the 25% and 75% quantiles vocabulary sizes. This provides a more intuitive way of understanding the patterns of vocabulary acquisition across age groups, as represented by the red dots, as well as the variability in typical vocabulary size measurements (grey error bars). VST norms are available for the following L1s: English, Table H.1 (page 217) and Figure H.1 (page 216); Russian, Table H.2 (page 219) and Figure H.2 (page 218); Chinese (Mandarin), Table H.3 (page 221) and Figure H.3 (page 220); Japanese, Table H.4 (page 223) and Figure H.4 (page 222); Arabic, Table H.5 (page 225) and Figure H.5 (page 224); Vietnamese, Table H.6 (page 227) and Figure H.6 (page 226); Indonesian, Table H.7 (page 229) and Figure H.7 (page 228); Spanish, Table H.8 (page 231) and Figure H.8 (page 230); German, Table H.9 (page 233) and Figure H.9 (page 232); and Korean, Table H.10 (page 235) and Figure H.10 (page 234).

The norms for English follow the expected pattern: continuously increasing vocabulary sizes throughout childhood and then beginning to taper off in the 20–30 age range. Vocabulary size for L1 English users becomes more or less flat around age 30 but there is still some slight increase throughout midlife. There is not enough data to confirm
what happens later in life, but the L1 English norms are suggestive of a very subtle, slow decline after around age 65.

It is interesting to note that this pattern appears as well for L2 English users whose native languages are Russian, Chinese (Mandarin), or Spanish. This might occur for L2 English users whose native languages are German or Korean as well, but the data are probably too sparse to make this conclusion. The upper bound of these norms is of course lower than that of L1 English users. L1 Russian and L1 Chinese users seem to attain a knowledge of around 9000 word families while L1 Spanish users appear to reach a level of just over 10,000 word families. A critical difference between these groups is that the age at which the increase on vocabulary sizes begin to taper off. For L1 Russian and L1 Spanish users, this seems to occur at roughly the same age as L1 English users, in the 20–30 age range.

For L1 Chinese users, however, the increase takes place about 10 years later. Why this is the case is unknown, but one possibility is the linguistic distance between English and Chinese. The plot of the L1 Japanese user English vocabulary sizes may support this idea. Although the L1 Japanese users do not exhibit the same logarithmic growth profile, the mean vocabulary size norms for most of the age groups are quite close for both L1 Japanese users and L1 Chinese users.

In sharp contrast, both L1 Arabic users and L1 Vietnamese users appear to have relatively flat vocabulary size profiles for all age groups, never really reaching the vocabulary sizes attained by other language groups. As with the vocabulary sizes of L1 Chinese users, it is difficult to explain why this should be the case. Part of the explanation might also lie in the distance between English and these languages, but language policy must play a role as well. Vietnam, for instance, previously emphasised the teaching of French and Russian and only recently turned attention to the teaching of English.

6.7 Limitations

Norms are only as reliable as the sample from which they are derived. Establishing reliable norms is basically a matter of collecting sufficient data from enough members of a population which represent a range of characteristics of interest. In this case, there are two variables which are easily measured and routinely reported in applied linguistics studies: participant L1 and age. While measurement is simple, collecting a representative sample is difficult. The L1s of internet users are strongly biased towards English for historical reasons. The level of technological advancement of the countries where a language is spoken is also important. Age is also a challenging variable when parti-
cipants’ self-selection is related to the level of one’s technical savvy. Older generations are probably less likely to be motivated to measure their vocabulary size online so those that do appear in the data are much less likely to be representative of their age group.

The vast majority of the test results reported in this chapter came from anonymous visitors to the VocabularySize.com website who took the test under their own volition. The sampling frame is therefore undefined. In such a situation, it is natural to assume that these participants do not represent the population at large as their motivations for taking the test in the first place are likely to include a stronger interest in vocabulary acquisition or a history of greater than average effort in increasing their mental lexicon. Indeed, discussions about the results of the VST taken on the website tend to be found in online forums dedicated to learning languages. With such an obvious bias in the data, great caution must be applied when interpreting the results; they are likely to primarily represent dedicated language learners and those with naturally larger vocabularies.

No test is perfect. The process of piloting and establishing norms has the potential to improve a test, but trial by fire is perhaps the greatest opportunity to uncover flaws and limitations that may never come to light in a more sheltered environment. To facilitate unfiltered feedback from participants who took the VST, an open-ended question soliciting comments about the test appeared at the very end of the test, just before the results were revealed. The majority of completed tests do not contain any comments at all. Of those that do, the vast majority of the comments are positive and supportive, often indicating that the test was “fun” or “good”. One of the most positive comments was, “I teach English and would just like to congratulate you - it’s wonderful to have this resource online!”

Limitations also became apparent and would need to be addressed in future iterations of the test and in data collection when revising the norms presented in this chapter. A thematically categorised list of some indicative, unedited comments can be found in Appendix G (page 209). The range of comments make it clear that no test will be ideal for every participant in every situation. For L1 English users, for example, the use of the BNC as the source of the items gives a distinct impression of dialectal bias leaving some American English users feeling disadvantaged. Even for some L2 users, items appeared to be selected from either nonstandard English or even American English. While greater effort can be made to sample items from more dialectally balanced corpora, the disproportionate use of some words over others by groups of languages users is inevitable.

Another major issue with the items, particularly among adult L1 English users, is the oversimplified nature of the options. The most common complaint is that the options are

---

7e.g. http://how-to-learn-any-language.com/forum/
CHAPTER 6. MEASURING VOCABULARY SIZE

“vague”, leading to the perception that none of them would be the correct response. The options for MICROPHONE generate a significant number of comments by those who note that the putative correct response, “machine that makes sounds louder”, describes an amplifier, not a microphone. More technically correct options could be provided for the vast majority of the items, but the increase in the difficulty level of the options and time required to read the items would probably cause more harm than good. The simplistic options are part of the design goals (P. Nation & Beglar, 2007) which prevent the comprehension of the options from being a barrier to accurate vocabulary size estimation.

Difficulty level is also a common complaint, both in terms of being too easy and too difficult. This is understandable because the paper-based version of the VST can make no assumptions about a participant’s likely vocabulary size. Participants with relatively small vocabulary sizes face the prospect of responding to many items which represent unknown word families. Likewise, those with relatively large vocabulary sizes spend the majority of the test responding to items which represent extremely well-known word families. Given this situation, it is likely that most participants will spend most of their time responding to items which add little to the accuracy of their vocabulary size determination.

The online version of the VST created for this study, and the norms established here, can immediately address most of these complaints. The average 18-year-old L1 English speaker, for example, will usually know between 10,000 and 17,000 word families (see Table H.1, page 217). At this level, testing the knowledge of high frequency items such as MICROPHONE does not improve the accuracy of the test results because it would be exceedingly rare for a participant with that vocabulary size to not know MICROPHONE. More useful information could be gathered by removing most items representing word families below and above the normal range for adult L1 English participants and increasing the items representing word families in the expected range indicated by the norms. By simply determining at the outset a participant’s age and L1, the online version of the VST could adjust the items presented to the participant to maximise the useful information gained from them. Dialectal differences could similarly be addressed by selecting items which best represent a participant’s preferred variety of English.

While most comments probably do contain a kernel of truth which can help inform future revisions, the unsolicited advice of one anonymous participant puts the comments into perspective: “[c]omments are dangerously slippery things.” Nevertheless, by converting the original VST to a dynamic, web-based resource backed by a database, many of the issues spontaneously identified by participants can be addressed through simple changes to the code that runs on the HTTP server. Different versions can be created as
well to facilitate data collection which can then be used to evaluate the effectiveness of the changes.

6.8 Conclusion

This chapter has established a large set of VST norms for a wide range ages for both L1 English native users and L2 English users. Measuring vocabulary size is a relatively rapid means of estimating language proficiency, so making the VST available online greatly increases the test’s appeal as a standardised measure of language proficiency for language practitioners and in future research.

As noted in Bernhardt’s (2005) model of L2 reading comprehension, language proficiency is expected to explain around 30% of L2 reading comprehension variance. While this is not enough to confidently predict reading comprehension to a significant degree, it is a substantial amount. Increasing the speed at which a measure of L2 proficiency, such as vocabulary size, is measured can free up time and resources to investigate the other 70% of L2 reading comprehension variance. Improving the accuracy of the measurement may result in an increase in explained reading comprehension variance as well.
Chapter 7

Conclusion

The process of science is self-correcting because humans are inherently fallible. Ironically, science has done an exceptional job of documenting how human perceptions, emotions, and cognition diverge from the dispassionate, stoic ideals of science. But human scientists will never achieve the idyllic qualities embedded in the scientific method, which is why replication is so important. It forces many eyes to investigate the same issue, eventually making each individual’s deeply flawed perceptions collectively shallow.¹

In reality, it takes time for this process to work. Even when research is replicated, interpretation and acceptance still filters through human minds, which can lead to cargo cult science, a term Feynman (1997) used in a commencement address given to graduates of California Institute of Technology, commonly known as Caltech, in 1974. In that speech, Feynman recounted several historical examples of cargo cult science such as the amount of time it took to arrive at the presently agreed value of the charge of an electron despite the experiment to determine the charge being purely based on measurement and devoid of any interpretation. Another example was the lack of interest in, and citations of, a psychology researcher who systematically eliminated plausible reasons why rats were able to negotiate a maze in ways which should not have been possible. Through the process of elimination, that researcher discovered a flaw in the assumptions about

¹This statement is a poor rephrasing of Linus’s Law, coined by Raymond (2001) and named in honour of Linus Torvalds, the creator of the Linux kernel. Its most basic form states that “given enough eyeballs, all bugs are shallow.” (p. 30) The law applies within the realm of computer programming, and specifically to the free- and open-source software movements, where flaws in a programme’s source code are easily detected and repaired if enough people are involved in the writing and testing processes of software development. The concept equally applies to any system where repetition by others is expected to reduce error.
how the mazes were routinely constructed for those types of experiment. Yet, according to Feynman, subsequent researchers have paid little heed to the discovery, repeating the mistakes despite researchers having access to the knowledge needed to avoid them.

According to Burridge (2001) the term cargo cult, without Feynman’s (1997) addition of “science”, refers to groups, primarily in Micronesia, which came into contact with much more technologically advanced groups, typically Western cultures. Due to the extreme difference in technological sophistication, many aspects of the advanced group were misunderstood at a fundamental level. In some instances, the less advanced group would begin to mimic superficial aspects of the more advanced group with the incorrect expectation that doing so would lead to the same benefits. In particular, many of the less advanced groups desired the “cargo” that the advanced groups regularly received. The lack of perspective beyond the local community and ignorance of trade and supply chains led to “cargo cults” which believed that the advanced groups received the cargo for free. By mimicking the objects and actions of the advanced group, the cargo cult would expect cargo to be delivered to them as well, which, of course, it never was.

Applying the concept to the process of doing science, Feynman suggests that researchers can fall into a similar trap. Humans have a natural tendency to succumb to inherent flaws in human perception and cognition which lead to confirmation bias, an unconscious process which overemphasises confirmatory evidence while systematically dismissing evidence or explanations which could refute a pre-existing belief. Confirmation bias makes it difficult to dismiss a currently held belief even in the face of overwhelming contradictory evidence. Without a deep level of understanding of how science and statistics work, researchers may mimic the process of data collection and analysis on only a superficial level. Unlike the cargo cults of Micronesia, researchers may even reap the reward of putative support for a hypothesis. But confirmation bias, and a lack of understanding of statistics and the scientific process, may prevent the researcher from investigating the data and analyses at a deeper level.

Feynman’s (1997) commencement address mentioning cargo cult science is apropos to mention here not simply because the concept may explain the development and appeal of the text coverage model of reading comprehension, but also because Feynman specifically mentions education research as an example of how researchers can fool themselves:

So I found things that even more people believe, such as that we have some knowledge of how to educate. There are big schools of reading methods and mathematics methods, and so forth, but if you notice, you’ll see the reading scores keep going down—or hardly going up—in spite of the fact that we
continually use these same people to improve the methods. [...] It ought to be looked into; how do they know that their method should work? [...] Yet these things are said to be scientific. We study them. And I think ordinary people with commonsense ideas are intimidated by this pseudoscience. A teacher who has some good idea of how to teach her children to read is forced by the school system to do it some other way—or is even fooled by the school system into thinking that her method is not necessarily a good one. [...] So we really ought to look into theories that don’t work, and science that isn’t science. (Feynman, 1997, p. 310)

Is it possible that research supporting the text coverage model of reading comprehension has fallen into the trap of cargo cult science without realising it? Or for that matter, how many of the currently espoused theories, methods, and ideologies in the field of applied linguistics in general are supported by multiple experiments which try to falsify the theory and fail? How many of those experiments have been repeated?

The social sciences, which subsume applied linguistics, have been undergoing a revolution in the way research is undertaken, reported, and interpreted. After decades of accumulating data, meta-analyses and retrospection are starting to indicate that many beliefs and theories are not as well supported as once thought. The discovery of many published studies in which researchers misused or misunderstood, both intentionally and naively, the scientific process (Open Science Collaboration, 2015) does not bode well. In the relatively new field of applied linguistics, the dearth of replication studies, especially of those which form the foundations of other studies, makes it highly probable that much, if not most, of the current state of knowledge is incorrect.

To address these issues, academic journals are beginning to require more transparency in reporting results (Eich, 2014). These requirements include a shift in statistical analyses away from simply reporting p values in null-hypothesis significance testing towards confidence intervals and effect sizes. Researchers are being encouraged, or even required, to more thoroughly describe any data manipulation or deletions prior to performing statistical tests. Data repositories are likewise gaining in popularity as a way of sharing original data, the code necessary to reproduce the analyses, and “fingerprints” of the data to ensure data integrity (Crosas, 2011).

Pre-registration of research is another increasingly common requirement. It requires researchers to publicly specify an experiment’s design and proposed analysis in advance of data collection (Open Science Collaboration, 2015). This requirement not
only addresses the Rosenthal’s (1979) File Drawer Problem, but also O’Boyle, Banks, and Gonzalez-Mulé’s (2014) related publication bias problem known as the *Chrysalis Effect*. Much like a beautiful butterfly emerges from its chrysalis after spending some time maturing in its protective environment, the Chrysalis Effect in research occurs when questionable research practices, such as *post-hoc* data manipulation, hypothesis revision, and hypothesis omission, occur after data are collected in order to improve the odds of publication (O’Boyle et al., 2014). It is a more active, intentional activity than the File Drawer Problem and can be seen clearly in the change in hypotheses, both revision and omission, when comparing Hu (1999) and Hu and Nation (2000).2

7.1 Contribution to the field

It is riskier to assert that a widely accepted model may in fact be wrong than to propose a new one, but the research described in this thesis strongly suggests that the text coverage model of reading comprehension may not be as skillful as previously thought. This research joins a new wave in the social sciences where highly regarded studies are subjected to replication and often found to be much less compelling than originally thought (e.g. Open Science Collaboration, 2015). Although the text coverage model of reading comprehension is intuitively appealing due to its simplicity and the ease at which it can be determined, based on the data, analysis, and critique in this thesis, it is time to openly question whether text coverage is actually just a gross oversimplification of the reading process. It is difficult to dismiss the fact that the amount of experimental evidence in support of its claims is underwhelming.

Certainly vocabulary is an important part of reading comprehension, and it makes intuitive sense that a reader would need to know the vast majority of words in a text for some purposes and degrees of comprehension, but not all. For example, having a low text coverage would mean greater opportunities to acquire vocabulary but the cost would be increased time in reaching certain levels of comprehension. The view of reading and comprehension implicit in the text coverage model of reading comprehension is idealised and narrowly constructed as some activity similar to extensive reading or Carver’s (2002) rauding. It is an image of language learners optimally matched to texts, reading speed and rate of unknown vocabulary encounters carefully controlled, and both text comprehension and incidental vocabulary learning potential fully optimised. This is a Utopian ideal which does not exist in the reality of piecemeal, incremental, messy

---

2Details on the Chrysalis Effect in Hu and Nation (2000) can be found in section 3.1 (page 59).
language acquisition and inherently wide confidence intervals. Its ecological validity is close to zero. What, then, is the contribution of the research in this thesis if the text coverage model of reading comprehension has not been improved? The primary purpose of research is not necessarily to invent new ideas or discover new phenomena. The bulk of scientific endeavour is spent trying to disprove existing ideas and theories in order to find those rarefied few which withstand all the attempts. It is only then that ideas begin to transition into the realm of truth. Although the proposal for this thesis was brimming with lofty goals of improving the power and precision of the text coverage model from a variety of converging approaches, the contribution to the field must ultimately be the discovery that text coverage as an accurate and reliable predictor of a considerable portion of reading comprehension is not as robust or substantial as once thought. This unexpected, and frankly disappointing, conclusion is supported empirically, through a faithful replication and expansion study which showed an attenuated effect and surprising regression coefficients.

This thesis also documented some of the primary ways that the recent line of text coverage research has lost track of its own theoretical foundations. The original meaning of threshold, one that was inherent in a reader, has been lost and replaced with an overly simplistic, atheoretical version that exists only at the interface of a reader and a text. The trajectory of the misunderstandings and oversimplifications has been documented to assist in avoiding similar problems in future. This may allow interested researchers to return to the original question of a L2 proficiency threshold and attempt to verify those predictions. Additionally, the sincere critique of the model itself gave further support to the conclusion that text coverage is not a tenable model of reading comprehension. It does not work in practice to an appreciable degree of accuracy and it will not work in theory.

By collecting tens of millions of data points from $\approx 1.31 \times 10^5$ L1 and L2 English users around the world, this research has aimed at a related target which holds much greater promise. Refinements to the VST (P. Nation & Beglar, 2007), and the creation of an online resource designed to save countless hours that would otherwise be spent manually marking the paper-based versions of the test, will enable practitioners and researchers to begin to standardise at least one factor involved in reading comprehension. This standardisation, which is common in psychology, can help push reading research in applied linguistics forward. The online VST provides a fast, inexpensive, and standardised method which, if used consistently, can be compared across studies to make them amenable to meta-analysis.
CHAPTER 7. CONCLUSION

7.2 Limitations

The results of any study could always be a fluke. Errors in data entry and analysis are virtually guaranteed. Omissions and oversights are rarely noticed, even long after publication. The critical value for significance at \( p \leq 0.05 \) guarantees that 5% of all results will be false. This study is no exception. In addition to the explicitly stated limitations below and despite due care on the part of the author, there is a high probability that much of the research presented in this thesis is factually incorrect.

Each chapter separately discussed limitations specific to the study, but there is still one issue which relates to both the replications and critique of the text coverage model of reading comprehension. It is possible that text coverage could still be shown to explain a substantial portion reading comprehension variance. There are several scenarios where this might be true:

- Thai participants were chosen not only because of the author’s experience living and working in Thailand, but also due to the linguistic distance between Thai and English. This distance could have had the opposite effect. Rather than accentuating a potential threshold, it could have attenuated it.

- The average age and proficiency levels of the Thai participants were lower than the participants in Hu and Nation (2000). It is possible that the effect of text coverage only comes into play once the language proficiency threshold has been crossed. At that point, L1 reading skills theoretically become available for use in L2 reading. Some of those skills likely include strategies for dealing with unknown words. Below the language proficiency threshold, the effects of unknown words on reading comprehension may not be measurable because the skills needed to deal with them are not available in the first place.

7.3 Future research

There is little use in further pursuing the text coverage model of reading comprehension. Researchers who choose to ignore this advice should be prepared to demonstrate that text coverage can explain more reading comprehension variance than existing readability formulae. One obvious challenge to doing so is the level of syntactic complexity in a text is probably strongly correlated with its lexical profile. As a text’s lexical profile becomes richer, i.e. the range of word types increases and average frequency of those types decreases, the average level of syntactic complexity of the text is also likely to
increase. Focusing on text coverage alone ignores this important relationship and the degree of readability that syntax brings to a text. Almost all readability formulae already account for both syntactic and lexical complexity.

The real world applications of the text coverage model, matching readers to texts and predicting reading comprehension, would need to be shown to be demonstrably easier and more successful than using existing readability formulae. It is unclear, however, how the actual proportion of unknown tokens in a non-experimental text could be calculated given existing technologies. Measuring the size of a reader’s mental lexicon is imprecise, yet determining text coverage would require precision in measuring not only the size, but also the actual contents. It appears to be impossible. It is currently possible to measure L2 proficiency and L1 reading skills sufficiently accurately enough to use them in multivariate readability formulae which yield useful reading comprehension predictions. The text coverage model requires a much greater level of precision to make more useful predictions.

Researchers would also need to explain any differential effects at the word level, such as the when key words are unknown. For example, a domain expert may experience very low text coverage overall, but very high text coverage of key vocabulary allowing the reader to fill in gaps in both syntactic and vocabulary. Meanwhile, another reader may experience high text coverage but the few unknown words may all be key vocabulary, thus making reading comprehension problematic.

If the critical review of the text coverage model, particularly the statistical impossibility of substituting cumulative text coverage and vocabulary size has failed to shake the theoretical foundations of text coverage research, then the following issues could be further investigated:3

- The variance of text coverage across genres with respect to the reader could be measured. This would involve increasing the number of texts read by each participant such as in Yoon (2011). It is inadvisable to generalise from small samples, especially when \( N = 1 \) as in many text coverage studies. If it turns out that the effect of text coverage is not invariant across genres, then ideal text coverage levels would need to be specified for each genre.

- To remain true to the theoretical framework on which Laufer and Sim’s (1985) study was based, investigations into the text coverage model could be made to

---

3These ideas were not pursued in this thesis because the results of such research would most likely lead to further nails in the coffin of text coverage rather than open up avenues for renewed confidence in the model.
show that the amount of reading comprehension variance explained by L1 ability is not only related to text coverage, but also that it is monotonic and nonlinear. A text coverage threshold, should one exist, would manifest itself by mediating a relationship between some L1 ability and L2 reading comprehension which would be more reliably modelled by a multiparametric logistic regression than a linear regression. The question to guide this research would be: at what text coverage does one’s L1 reading abilities become available for use in L2 reading comprehension?

7.4 Implications

Turning back to one of the grandparents of text coverage studies, M. West (1968) searched for a “minimum adequate” (p. 205), the least amount of L2 proficiency needed to hold a basic conversation, for more than 40 years, but never found it. This is a noble goal. Indeed, the goal seems to be shared by researchers whose corpus-driven studies result in vocabulary size recommendations for learners and are based on the text coverage model. But attempting calculate a vocabulary size needed to read a type of text through the middle man of text coverage is backwards.

A wealth of studies exist which show strong correlations between vocabulary knowledge and reading comprehension (see R. Anderson & Freebody, 1981) without the need for text coverage. The most promising path for actually determining a minimum adequate vocabulary size, similar to what M. West sought, is to improve the accuracy at which L2 abilities can be measured directly. More accurate vocabulary size measures, ones that measure depth, breadth, and fluency of access, will give researchers and practitioners valid insights about the learners themselves. Besides, even if the minimum adequate vocabulary size recommendations derived from text coverage turn out to be reasonably accurate, the need for better, faster, more accurate tests of vocabulary size to determine when a learner has reached that goal still remains! Recommendations mean little if there are no good ways of determining when they have been met. The future of the quest for a minimum adequate vocabulary size, therefore, lies primarily in measuring the vocabulary size of readers, not in the overlap between a reader’s assumed mental lexicon and the lexical profile of a text.
Appendix A

Changes to *The Escaped Lunatic*
Table A.1: Changes to the original, unmodified *The Escaped Lunatic* used in Hu and Nation (2000).

<table>
<thead>
<tr>
<th>Original</th>
<th>Simplification</th>
<th>Pseudoword</th>
</tr>
</thead>
<tbody>
<tr>
<td>anxious</td>
<td>stergic</td>
<td></td>
</tr>
<tr>
<td>anxiously</td>
<td>stergically</td>
<td></td>
</tr>
<tr>
<td>bit</td>
<td>kiney</td>
<td></td>
</tr>
<tr>
<td>brain</td>
<td>clion</td>
<td></td>
</tr>
<tr>
<td>burst</td>
<td>pluched</td>
<td></td>
</tr>
<tr>
<td>chair</td>
<td>bristy</td>
<td></td>
</tr>
<tr>
<td>collect</td>
<td>plualy</td>
<td></td>
</tr>
<tr>
<td>crash</td>
<td>drach</td>
<td></td>
</tr>
<tr>
<td>crashed</td>
<td>drached</td>
<td></td>
</tr>
<tr>
<td>crashes</td>
<td>draches</td>
<td></td>
</tr>
<tr>
<td>dining</td>
<td>morting</td>
<td></td>
</tr>
<tr>
<td>dinner</td>
<td>pestance</td>
<td></td>
</tr>
<tr>
<td>firm</td>
<td>pruim</td>
<td></td>
</tr>
<tr>
<td>flashed</td>
<td>scrizzed</td>
<td></td>
</tr>
<tr>
<td>football</td>
<td>cozinet</td>
<td></td>
</tr>
<tr>
<td>governor</td>
<td>phantroper</td>
<td></td>
</tr>
<tr>
<td>joke</td>
<td>shreng</td>
<td></td>
</tr>
<tr>
<td>knocked</td>
<td>cranged</td>
<td></td>
</tr>
<tr>
<td>knocker</td>
<td>cranger</td>
<td></td>
</tr>
<tr>
<td>knocking</td>
<td>cranging</td>
<td></td>
</tr>
<tr>
<td>leg</td>
<td>dourn</td>
<td></td>
</tr>
<tr>
<td>locked</td>
<td>rounced</td>
<td></td>
</tr>
<tr>
<td>madman</td>
<td>vesbian</td>
<td></td>
</tr>
<tr>
<td>meanwhile</td>
<td>berously</td>
<td></td>
</tr>
<tr>
<td>obviously</td>
<td>emartically</td>
<td></td>
</tr>
<tr>
<td>pause</td>
<td>fabut</td>
<td></td>
</tr>
<tr>
<td>peculiar</td>
<td>utermous</td>
<td></td>
</tr>
<tr>
<td>railway</td>
<td>cubulare</td>
<td></td>
</tr>
<tr>
<td>relieved</td>
<td>lerized</td>
<td></td>
</tr>
<tr>
<td>rushed</td>
<td>blerged</td>
<td></td>
</tr>
<tr>
<td>searching</td>
<td>swanding</td>
<td></td>
</tr>
<tr>
<td>shut</td>
<td>garnaged</td>
<td></td>
</tr>
<tr>
<td>slippery</td>
<td>adinned</td>
<td></td>
</tr>
<tr>
<td>suspicion</td>
<td>farsacion</td>
<td></td>
</tr>
<tr>
<td>telephone</td>
<td>tangerite</td>
<td></td>
</tr>
<tr>
<td>telephoned</td>
<td>tangerited</td>
<td></td>
</tr>
<tr>
<td>thank</td>
<td>feund</td>
<td></td>
</tr>
<tr>
<td>warn</td>
<td>glet</td>
<td></td>
</tr>
<tr>
<td>washed</td>
<td>pemured</td>
<td></td>
</tr>
<tr>
<td>wires</td>
<td>pundries</td>
<td></td>
</tr>
<tr>
<td>yard</td>
<td>rander</td>
<td></td>
</tr>
<tr>
<td>yards</td>
<td>randers</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

VST localised for Thai
First 1000

1. see: They saw it.
   (a) ตด
   (b) รอคอย
   (c) เหน
   (d) เรมตน
   (e) ไมทราบ

2. time: They have a lot of time.
   (a) เงน
   (b) อาหาร
   (c) เวลา
   (d) เพอนๆ
   (e) ไมทราบ

3. period: It was a difficult period.
   (a) คำถาม
   (b) ชว
g
   (c) งาน
   (d) หนังสือ
   (e) ไมทราบ

4. figure: Is this the right figure?
   (a) คำตอบ
   (b) สถานที่
   (c) เวลา
   (d) ดำดเสя
   (e) ไมทราบ

5. poor: We are poor.
   (a)ยากจน
   (b) มีความสุข
   (c) สนใจ
   (d) ง่าย
d
   (e) ไมทราบ

6. drive: He drives fast.
   (a) เหมือน
   (b) เรียนรู้
   (c) การ
   (d) รถ
   (e) ไมทราบ

7. jump: She tried to jump.
   (a) ลอยน้ำ
   (b) กระโดด
   (c) จอดแล้ว
   (d) ซิด
   (e) ไมทราบ

8. shoe: Where is your shoe?
   (a) ผ้า
   (b) กระเป๋า
   (c) ปากกา
   (d) รองเท้า
   (e) ไมทราบ

9. standard: Her standards are very high.
   (a) เหนือ
   (b) มีผลเสถ
   (c) ราย
   (d) กระบะเสถ
   (e) ไมทราบ

10. basis: I don't understand the basis.
    (a) ของขวัญ
    (b) ผล
    (c) ความ
    (d) คำ
    (e) ไมทราบ

Second 1000

1. maintain: Can they maintain it?
   (a) ปรับปรุง
   (b) เท่า
   (c) รักษา
   (d) ได้รับ
   (e) ไมทราบ

2. shoe: Where is your shoe?
   (a) ผ้า
   (b) กระเป๋า
   (c) ปากกา
   (d) รองเท้า
   (e) ไมทราบ

3. upset: I am upset.
   (a) เหนือ
   (b) มีผลเสถ
   (c) ราย
   (d) กระบะเสถ
   (e) ไมทราบ

4. drawer: The drawer was empty.
   (a) ลิ้น
   (b) ถูด
   (c) ผิว
   (d) ของ
   (e) ไมทราบ

5. patience: He has no patience.
   (a) ความอดทน
   (b) เวลา
   (c) ความ
   (d) ความเดิน
   (e) ไมทราบ

6. nil: His mark for that question was nil.
   (a) ผล
   (b) ข้อ
   (c) ความ
   (d) ความ
   (e) ไมทราบ

7. pub: They went to the pub.
   (a) บาร์
   (b) ธนาคาร
   (c) ร้าน
   (d) สรรพ
   (e) ไมทราบ

8. circle: Make a circle.
   (a) ภาพโครงสร้าง
   (b) ความ
   (c) วงกลม
   (d) รู
   (e) ไมทราบ

9. microphone: Please use the microphone.
   (a) ที่ออก
   (b) เครื่องขยายเสถ
   (c) กล้อง
   (d) โทรศ
   (e) ไมทราบ

10. pro: He's a pro.
    (a) นักเสี่ยญ
    (b) นักกีฬา
    (c) ผู้เข้าเ
    (d) ตนเอง
    (e) ไมทราบ
Third 1000
1. soldier: He is a soldier.
   (a) นกธ(รกจ
   (b) ทหาร
   (c) ชางต%เหลก
   (d) นกศ	กษา
   (e) ไมทราบ
2. restore: It has been restored.
   (a) พ+ดซ
   (b) แจกใหม
   (c) ลดราคา
   (d) ฟ*นฟ+
   (e) ไมทราบ
3. jug: He was holding a jug.
   (a) เหยอก
   (b) การอภปราย
   (c) หมวก
   (d) ล+กระเบด
   (e) ไมทราบ
4. scrub: He is scrubbing it.
   (a) สลก
   (b) ขอมแซม
   (c) ถ
   (d) ร่างกาย
   (e) ไมทราบ
5. dinosaur: The children were pretending to be dinosaurs.
   (a) โจรสลด
   (b) เทวดา
   (c) มงกร
   (d) สตว1ขนาดใหญ่มากท%ส+ญ
   (e) ไมทราบ
6. rove: He couldn't stop roving.
   (a) ดมจด
   (b) ทองเท%ยวไป
   (c) ครวญเพลง
   (d) งานหนก
   (e) ไมทราบ
7. pave: It was paved.
   (a) ปดก*น
   (b) แบง
   (c) เคลอบทอง
   (d) ป+พ*น
   (e) ไมทราบ
8. dash: They dashed over it.
   (a) โผเขาหา
   (b) ออยอง
   (c) ตอส+
   (d) ช!าเลอง
   (e) ไมทราบ
9. candid: Please be candid.
   (a) ชอบบั้ย
   (b) ทอง
   (c) สตวเลียง
   (d) นิวผาแม่เอม
   (e) ไมทราบ
10. lonesome: He felt lonesome.
    (a) อกตญญ+
    (b) หมดแรง
    (c) หงอยเหงา
    (d) asurement
    (e) ไมทราบ

Fourth 1000
1. compound: They made a new compound.
   (a) ขอตกลง
   (b) สารประกอบ
   (c) สมาคม
   (d) การคาดเดา
   (e) ไมทราบ
2. latter: I agree with the latter.
   (a) ผาคล(มหว
   (b) ทอง
   (c) สตวเลียง
   (d) นิวผาแม่เอม
   (e) ไมทราบ
3. quiz: We made a quiz.
   (a) กระบอกบรรจ(ล+กธน+
   (b) ความผดพลาดอยางราย
   (c) การสอบ
   (d) กรงนก
   (e) ไมทราบ
4. tummy: Look at my tummy.
   (a) ผาคล(มหว
   (b) ทอง
   (c) สตวเลียง
   (d) นิวผาแม่เอม
   (e) ไมทราบ
5. input: We need more input.
   (a) ขอม+ล
   (b) ล+กจาง
   (c) เงน
   (d) ไมทราบ
   (e) ไมทราบ
6. crab: Do you like crabs?
   (a) ป+
   (b) ขนมปงกรอบ
   (c) คอปกเส
   (d) จ*งหร%ด
   (e) ไมทราบ
7. vocabulary: You will need more vocabulary.
   (a) เงน
   (b) ทกษะ
   (c) ค!าศ
   (d) ปน
   (e) ไมทราบ
8. remedy: We found a good remedy.
   (a) สวนสาธารณะ
   (b) การรกษา
   (c) ต!ารบ
   (d) ลำดับขยันตอนท%แนนอน
   (e) ไมทราบ
9. allege: They alleged it.
   (a) กลาวหา
   (b) ขโมยความคดผ+อน
   (c) พส+จน1
   (d) โตแยง
   (e) ไมทราบ
10. alge: They alleged it.
    (a) กล่าวหา
    (b) ขโมยความคดผ+อน
    (c) พส+จน1
    (d) โตแยง
    (e) ไมทราบ
Fifth 1000

1. **deficit**: The company had a large deficit.
   (a) เงินสะสม
   (b) ขาด
   (c) ขาด
   (d) ขาด
   (e) ไม่ทราบ

2. **weep**: He wept.
   (a) จบการศึกษา
   (b) ร้องไห้
   (c) ร้องไห้
   (d) ร้องไห้
   (e) ไม่ทราบ

3. **nun**: We saw a nun.
   (a) นUNS
   (b) บอด
   (c) บอด
   (d) บอด
   (e) ไม่ทราบ

4. **haunt**: The house is haunted.
   (a) ตกแต่ง
   (b) มีผี
   (c) มีผี
   (d) ผี
   (e) ไม่ทราบ

5. **compost**: We need some compost.
   (a) การสนับสนุน
   (b) ความสะดวกสบาย
   (c) ปน
   (d) ปน
   (e) ไม่ทราบ

6. **cube**: I need one more cube.
   (a) ตัว
   (b) ลิข
   (c) ลิข
   (d) ลิข
   (e) ไม่ทราบ

7. **miniature**: It is a miniature.
   (a) สีที่มีขนาดเล็กมาก
   (b) กระเบื้อง
   (c) กระเบื้อง
   (d) กระเบื้อง
   (e) ไม่ทราบ

Sixth 1000

1. **devious**: Your plans are devious.
   (a) หลอก
   (b) หลอก
   (c) หลอก
   (d) หลอก
   (e) ไม่ทราบ

2. **premier**: The premier spoke for an hour.
   (a) วิทยากร
   (b) แล้ว
   (c) แล้ว
   (d) แล้ว
   (e) ไม่ทราบ

3. **butler**: They have a butler.
   (a) แค้น
   (b) แฟ้ม
   (c) แฟ้ม
   (d) แฟ้ม
   (e) ไม่ทราบ

4. **accessory**: They gave us some accessories.
   (a) ตัว
   (b) ตัว
   (c) ตัว
   (d) ตัว
   (e) ไม่ทราบ

5. **threshold**: They raised the threshold.
   (a) ธง
   (b) เล่น
   (c) เล่น
   (d) เล่น
   (e) ไม่ทราบ

6. **thesis**: She has completed her thesis.
   (a) วิทยานิพนธ์
   (b) วิทยานิพนธ์
   (c) วิทยานิพนธ์
   (d) วิทยานิพนธ์
   (e) ไม่ทราบ

7. **strangle**: He strangled her.
   (a) ปล้น
   (b) ให้
   (c) ให้
   (d) ให้
   (e) ไม่ทราบ

8. **cavalier**: He treated her in a cavalier manner.
   (a) อย่างสุภาพ
   (b) อย่างสุภาพ
   (c) อย่างสุภาพ
   (d) ไม่ทราบ
   (e) ไม่ทราบ

9. **malign**: His malign influence is still felt.
   (a) ราย
   (b) ผิด
   (c) ผิด
   (d) ผิด
   (e) ไม่ทราบ

10. **veer**: The car veered.
    (a) ตัด
    (b) ตัด
    (c) ตัด
    (d) ตัด
    (e) ไม่ทราบ
Seventh 1000
1. olive: We bought olives.
   (a) ผลมะกอก
   (b) ดอกไม้หอม
   (c) กางเกงวายนำ
   (d) จาบ
   (e) ไมทราบ
2. quilt: They made a quilt.
   (a) พนยกรรม
   (b) นตกรรมสัญญา
   (c) ผ้านวมคลุม
   (d) ผ้านวมกันหนาว
   (e) ไมทราบ
3. stealth: They did it by stealth.
   (a) การใช้เงินมหาศาล
   (b) การข่มขู่
   (c) การทั่วถึง
   (d) การเพลิดเพลิน
   (e) ไมทราบ
4. shudder: The boy shuddered.
   (a) กระซับ
   (b) เดินสะดุด
   (c) สวม
   (d) ตะโกน
   (e) ไมทราบ
5. bristle: The bristles are too hard.
   (a) คำถาม
   (b) ชานปง
   (c) เลี้ยงพับได้
   (d) พื้นรองเท้า
   (e) ไมทราบ
6. bloc: They have joined this bloc.
   (a) วงดนตรี
   (b) กลมมั่นคง
   (c) ทหารพราน
   (d) กลมประเทศ
   (e) ไมทราบ
7. demography: This book is about demography.
   (a) ภูมศาสตร์
   (b) สถิติ
   (c) วิชาที่เกี่ยวกับการเคลื่อนไหวของเหลว
   (d) ประชากรศาสตร์
   (e) ไมทราบ
8. azalea: This azalea is very pretty.
   (a) ตนไมชนดหนัง
   (b) ผาไหม
   (c) ผาสี
   (d) เปลอกหอยชนดหนัง
   (e) ไมทราบ
9. yoghurt: This yoghurt is disgusting.
   (a) โคลนด้าน
   (b) ฝา
   (c) นมเปรี้ยว
   (d) ผลิยางค์
   (e) ไมทราบ
10. bristle: The bristles are too hard.
   (a) คำถาม
   (b) ชานปง
   (c) เลี้ยงพับได้
   (d) พื้นรองเท้า
   (e) ไมทราบ

Eighth 1000
1. erratic: He was erratic.
   (a) สมบับสนธิ
   (b) เหตุผล
   (c) มีการกระทำ
   (d) ไม่แน่นอน
   (e) ไมทราบ
2. palette: He lost his palette.
   (a) ตะกร้าใส่ปลา
   (b) ความอบอุ่น
   (c) เพื่อนคู่
   (d) ทานผักสด
   (e) ไมทราบ
3. null: His influence was null.
   (a) ได้ผลดี
   (b) แปลงประโยคน
   (c) ไม่มี
   (d) ลอง
   (e) ไมทราบ
4. kindergarden: This is a good kindergarten.
   (a) งานอดิเรก
   (b) โรงเรียนอนุบาล
   (c) กระเป๋าเป้
   (d) ห้องสมุด
   (e) ไมทราบ
5. eclipse: There was an eclipse.
   (a) ลมแรง
   (b) สาดกระเด็น
   (c) การฆาตกรรม
   (d) การบดบังแสง
   (e) ไมทราบ
6. marrow: This is the marrow.
   (a) เครื่องราง
   (b) ปาก
   (c) ทะลุ
   (d) ลาย
   (e) ไมทราบ
7. locust: There were hundreds of locusts.
   (a) ดึกดัน
   (b) ราส
   (c) สนิท
   (d) ผลิยางค์
   (e) ไมทราบ
8. authentic: It is authentic.
   (a) ของแท้
   (b) กระชับ
   (c) วาดภาพ
   (d) แปลง
   (e) ไมทราบ
9. cabaret: We saw the cabaret.
   (a) จัดการแสดงเวที
   (b) การแสดงดนตรี
   (c) หนอนผ่าน
   (d) นางเงอก
   (e) ไมทราบ
10. mumble: He started to mumble.
    (a) ครึ่ง
    (b) สน
    (c) ลาหลง
    (d) พูดพูดพ่าน
    (e) ไมทราบ
Appendix C

Literacy Questionnaire
<table>
<thead>
<tr>
<th>แผน/ดิฉีวน...</th>
<th>การอ่านภาษาไทย</th>
<th>การอ่านภาษาอังกฤษ</th>
</tr>
</thead>
<tbody>
<tr>
<td>ชอบทำงานอย่างนี้มากกว่าการอ่าน</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>ชอบหนังสือไม่เยอะ</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>ไม่กลัวที่จะบอกใครว่า ไม่ชอบการอ่านเลย</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>มีหนังสือหลายเล่มไว้ในห้อง</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>ชอบอ่านตอนมีเวลาว่าง</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>รู้สึกดีแต่เรื่องหนังสือที่ได้อ่านไป</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>รักการอ่าน</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>ชอบอ่านหนังสือที่นักเขียนซื่อต่อกัน</td>
<td>เหนดวย</td>
<td>เหนดวย</td>
</tr>
<tr>
<td>ไม่มีหนังสือจากห้องสมดุล</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>ชอบอยู่บ้านอ่านหนังสือ</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>ไม่ค่อยอ่านหนังสือ ยกเว้นตอนที่ต้องอ่านเพื่อทำรายงาน</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>รู้สึกว่าการอ่านเสียเวลา</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>รู้สึกว่าการอ่านน่าเบื่อ</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>คิดว่าคนที่ชอบอ่านมาก เป็นคนแปลก</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>อ่านเพื่อนจากบัญชา</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>ชอบแก้ไขคนที่ชอบอ่านมาก</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>ชอบแบ่งหนังสือกับเพื่อนๆ</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>อยากให้ใครเล่าเรื่องให้ฟังมากกว่าที่ต้องอ่านเอง</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>เกลี้ยงการอ่าน</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>ยิ้มหน้างสือเมื่อไปห้องสมุด</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>ใช้เวลาอ่านหนังสือนาน</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>อ่านเพื่อจะได้เพิ่มความสนใจในสิ่งต่างๆ</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>อ่านหนังสือบ่อย</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>นำมาเศษใหม่จะนำไปใช้</td>
<td>เหนเด้ย</td>
<td>เหนเด้ย</td>
</tr>
<tr>
<td>ชอบได้รับหนังสือเป็นของขวัญ</td>
<td>เหนเด้ย</td>
<td>เ.uniform</td>
</tr>
</tbody>
</table>
ในช่วงปิดเทอมที่ผ่านมา ประโยคดังนี้: ประโยคภาษาไทย ประโยคภาษาอังกฤษ

| หนังสือพิมพ์ | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |
| บัตรยาส์ | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |
| หนังสือการ์ตูน | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |
| บทกวี | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |
| จดหมาย/email/ห้องสนทนากลาง | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |
| เว็บไซต์ | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |
| หนังสือจากเรื่องจริง | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |
| อื่นๆ (ระบุ: ____________) | อาทิตย์ละ ________ ชั่วโมง | อาทิตย์ละ ________ ชั่วโมง |

ข้อมูลส่วนตัว

อายุ ________
เพศ ________
ภาษาแม่ / ภาษาต้องการ (ระบุทั้งหมด)

ระยะเวลาเรียนภาษาอังกฤษ ________ ปี
ระยะเวลาอยู่หรือเที่ยวในประเทศที่ใช้ภาษาอังกฤษเป็นภาษาแม่ ________ เดือน
คะแนนล่าสุด (ถ้ามี):

TOEIC ________ เดือน/ปี ________
TOEFL ________ เดือน/ปี ________
IELTS ________ เดือน/ปี ________
Appendix D

Experimental Texts
I was in the middle of making some soup for dinner when the telephone rang. I washed my hands and went to answer it.

“Hello,” said an anxious voice when I had picked up the receiver, “could I speak to Mrs. Scott, please?”

“Speaking,” I answered.

“Mrs. Mary Scott?” the voice went on, even more anxiously.

“Yes,” I answered. “What’s the matter?” By now I was getting a bit frightened myself. Thoughts of my husband being in a railway crash, or of one of the children breaking his leg playing football in the school-yard raced through my mind.

“This is the Northfields Hospital,” said the voice. “One of our patients has escaped, and one of his friends heard him say that he was going to kill you. We thought we’d better warn you.”

“Oh, yes, thank you,” I answered and stopped to think very quickly. “What’s the name of this, er, patient?” I then went on.

There was a pause at the other end, as if someone was searching through a list. Then, “George Adams is his name,” came the reply. “He’s short and dark-haired and very dangerous.”

“But why does he want to, er, kill me? I’ve never even heard of him.”
A strange xxxx sound came from the telephone and it went dead. Thoughts of people cutting telephone wires before attacking a house flashed through my brain. Would a xxxx think of doing a thing like that?

I was still holding the receiver when there was a xxxx at the front-door. I dropped the receiver and raced to the door. I locked it just as the man outside was beginning to open the cover of the letter-box. Then I rushed to the back-door, knocking a small table over on my way and breaking the flower-bowl on it to pieces. The water made the floor smooth and I slipped and crashed into a chair, knocking that over too. However, I reached the back-door, locked it and then went around the house making sure all the windows were shut. They had no bars on them, so it would be easy to break one open.

Meanwhile, the telephone-operator, finding I had not put my telephone down and hearing the crashes and xxxx, thought there was a fight going on in my house and telephoned the police. In a few minutes, I heard heavy footsteps advancing up the drive and there was a firm sound on the knocker of the front door.

“Who’s that?” I said in a trembling voice.

“Police,” came the answer. “is there anything wrong, Lady?”

I thought fast. “Come round to the window so that I can see you,” I said. The heavy footsteps circled the garden, and two policemen, an officer and a xxxx, appeared at the dining-room window. I ran joyfully to the front door to let them in. The postman was standing a few yards off, looking at me in a way that struck me as unusual.

“Did you catch the madman?” I asked the officer.

“What madman?” he asked, clearly completely confused.

“Why, the one that was xxxx at my door. He escaped from the xxxx xxxx, and his name’s Adams. George Adams.”

“George Adams?” said the officer, even more confused than before. “But that’s the name of the Governor of the hospital.”

“And it was me that knocked at your door,” said the postman.

We looked at each other, the officer, the xxxx, the postman, and I. Fright, surprise, and suspicion followed each other across our faces. Then the police officer said, “Somebody must have been having a joke with you, lady, and a stupid joke at that.”

We all burst into relieved laughter.
ขอขอบคุณที่ให้ความอนุเคราะห์ในการทำแบบทดสอบ เราจะศึกษาผลกระทบของค่าพื้นที่ไม่รู้จักที่มีต่อความเข้าใจในเรื่องที่ถูกถาม แต่นั่นไม่ได้แปลความถูกต้องของค่าพื้นที่ที่ผูกตัวอยู่กับจุดที่ไม่รู้จักอย่างไรก็ตามคุณไม่จำเป็นต้องละเลยเมื่อเรื่องค่าพื้นที่เหล่านี้แลกไม่ให้คุณพยายามทำแบบทดสอบให้เข้าใจที่จะทำได้

ขั้นตอนที่ ๑ อ่านเนื้อเรื่องส่วนๆ คุณสามารถอยู่ที่เร็วที่ได้ หากที่คุณอ่านเสร็จแล้ว กรุณาตอบข้อตาม หลังจากนั้นผู้วิจัยจะมาวันเนื้อเรื่องในกระดาษไปและแจกกระดาษคำตอบเชุดที่ ๑ ให้กับคุณ

ขั้นตอนที่ ๒ อ่านเนื้อเรื่องให้ครบถ้วนข้อ กรุณาทำแบบทดสอบอย่างถูกต้อง ไม่ควรคาดเดาคำตอบ

• หากคุณจำไม่ได้ หรือไม่ให้ความเข้าใจในเนื้อเรื่อง ให้ดูข้อความเลือก (e) * รู้สึกว่าลำาในไปแล้วแต่กระไรจะถูกต้องไม่ได้

• หากคุณไม่ทราบคำตอบ กรุณาเลือก: (f) * ไม่ทราบ

ขั้นตอนที่ ๓ ที่คุณตอบคำถามเสร็จเรียบร้อยแล้ว กรุณาตอบข้อจากนั้นผู้วิจัยจะมาวันกระดาษคำตอบค่าตอบเก็บข้อมูลไปและแจกกระดาษคำตอบเชุดที่ ๒ ให้กับคุณ

The Escaped Selgian

I was in the middle of making some soup for dinner when the telephone rang. I washed my hands and went to answer it.

“Hello,” said an anxious voice when I had picked up the receiver, “could I speak to Mrs. Scott, please?”

“Speaking,” I answered.

“Mrs. Mary Scott?” the voice went on, even more anxiously.

“Yes,” I answered. “What’s the matter?” By now I was getting a bit frightened myself. Thoughts of my husband being in a railway crash, or of one of the children breaking his leg playing football in the school-yard raced through my mind.

“This is the Northfields Hospital,” said the voice. “One of our patients has escaped, and one of his friends heard him say that he was going to kill you. We thought we’d better warn you.”

“Oh, yes, thank you,” I answered and stopped to think very quickly. “What’s the name of this, er, patient?” I then went on.

There was a pause at the other end, as if someone was searching through a list. Then, “George Adams is his name,” came the reply. “He’s short and dark-haired and very dangerous.”

“But why does he want to, er, kill me? I’ve never even heard of him.”
A strange clairing sound came from the telephone and it went dead. Thoughts of people cutting telephone wires before attacking a house flashed through my brain. Would a selgian think of doing a thing like that?

I was still holding the receiver when there was a knock at the front-door. I dropped the receiver and raced to the door. I locked it just as the man outside was beginning to open the cover of the letter-box. Then I rushed to the back-door, knocking a small table over on my way and breaking the flower-bowl on it to pieces. The water made the floor smooth and I stranped and crashed into a chair, knocking that over too. However, I reached the back-door, locked it and then went around the house making sure all the windows were shut. They had no bars on them, so it would be easy to break one open.

Meanwhile, the telephone-operator, finding I had not put my telephone down and hearing the crashes and knocking, thought there was a fight going on in my house and telephoned the police. In a few minutes, I heard heavy footsteps advancing up the drive and there was a firm sound on the knocker of the front door.

"Who's that?" I said in a gramoning voice.

"Police," came the answer. "Is there anything wrong, Lady?"

I thought fast. "Come round to the window so that I can see you," I said. The heavy footsteps circled the quenery, and two policemen, an officer and a mirocan, appeared at the dining-room window. I ran joyfully to the front door to let them in. The postman was standing a few yards off, looking at me in a way that struck me as unusual.

"Did you catch the madman?" I asked the officer.

"What madman?" he asked, clearly completely poigafied.

"Why, the one that was knocking at my door. He escaped from the selgian tarden, and his name's Adams. George Adams."

"George Adams?" said the officer, even more poigafied than before. "But that's the name of the Governor of the hospital."

"And it was me that knocked at your door," said the postman.

We looked at each other, the officer, the mirocan, the postman, and I. Fright, surprise, and suspicion nurated each other across our faces. Then the police officer said, "Somebody must have been having a joke with you, lady, and a stupid joke at that."

We all burst into relieved laughter.
The Escaped Selgian

I was in the middle of making some rawe for dinner when the telephone rang. I washed my hands and went to answer it.

“Hello,” said an anxious voice when I had picked up the receiver, “could I speak to Mrs. Scott, please?”

“Speaking,” I answered.

“Mrs. Mary Scott?” the voice went on, even more anxiously.

“Yes,” I answered. “What’s the matter?” By now I was getting a bit soranced myself. Thoughts of my husband being in a railway crash, or of one of the children breaking his leg playing football in the school-yard raced through my mind.

“This is the Northfields Slivian,” said the voice. “One of our patients has escaped, and one of his sleards heard him say that he was going to kill you. We thought we’d better warn you.”

“Oh, yes, thank you,” I answered and stopped to dreed very quickly. “What’s the name of this, er, patient?” I then went on.

There was a pause at the other end, as if someone was searching through a list. Then, “George Adams is his name,” came the reply. “He’s short and dark-haired and very dangerous.”

“But why does he want to, er, kill me? I’ve never even heard of him.”
A strange clairing sound came from the telephone and it went dead. Thoughts of people cutting telephone wires before attacking a house flashed through my brain. Would a selgian think of doing a thing like that?

I was still holding the receiver when there was a creng at the front-door. I dropped the receiver and raced to the door. I locked it just as the man outside was beginning to open the cover of the letter-box. Then I rushed to the back-door, knocking a small table over on my way and higoning the flower-plurd on it to lacrams. The water made the nalut smooth and I stranped and crashed into a chair, knocking that over too. However, I reached the back-door, locked it and then went around the house making sure all the windows were shut. They had no bars on them, so it would be easy to break one open.

Meanwhile, the telephone-operator, finding I had not put my telephone down and hearing the crashes and crengs, thought there was a fight going on in my house and telephoned the police. In a few minutes, I heard heavy footsteps advancing up the drive and there was a firm sound on the knocker of the front door.

“Who’s that?” I said in a gramoning voice.

“Police,” came the answer. “is there anything wrong, Stragar?”

I thought fast. “Come round to the window so that I can see you,” I said. The heavy footsteps circled the quenery, and two policemen, an cheltian and a mirocan, appeared at the dining room window. I ran joyfully to the front door to let them in. The dripper was heaking a few yards off, looking at me in a way that struck me as unusual.

“Did you catch the madman?” I asked the cheltian.

“What madman?” he asked, clearly completely poigafied.

“Why, the one that was crenging at my door. He escaped from the selgian tarden, and his name’s Adams. George Adams.”

“George Adams?” said the cheltian, even more poigafied than before. “But that’s the name of the Governor of the slivian.”

“And it was me that knocked at your door,” said the dripper.

We looked at each other, the cheltian, the mirocan, the dripper, and I. Sorance, surprise, and suspicion nurated each other across our faces. Then the police officer said, “Somebody must have been having a joke with you, Stragar, and a prabic joke at that.”

We all burst into relieved laughter.
ขอขอบคุณที่ให้ความอนุเคราะห์ในการทำแบบทดสอบ เราประสงค์ที่จะศึกษาผลกระทบของคำศัพท์ที่ไม่รู้จักที่มีต่อความเข้าใจเนื้อเรื่องที่อ่าน ดังนั้นในแบบทดสอบนี้มีคำศัพท์บางคำที่คุณอาจไม่รู้จักอย่างไรก็ตามคุณไม่จำเป็นต้องตระหนักรู้คำศัพท์เหล่านี้ เพื่อให้คุณพยายามอ่านเนื้อเรื่องให้เข้าใจที่จะทำได้

ขั้นตอนที่ ๑ อ่านเนื้อเรื่องส่วนๆ คุณสามารถอ่านสิ่งที่ให้ทั้งที่คุณอ่านเสร็จแล้ว กรุณาตอบเนื้อเรื่องในกลับไปและจะแจกกระดาษคำถามชุดที่ ๒ ให้กับคุณ

ขั้นตอนที่ ๒ ตอบคำถามให้ครบถ้วน ถ้าคำถามบางคำถามคุณตอบไม่ได้ กรุณาเลือก: (e) * รู้สึกว่าอ่านไม่ได้ แต่ต้องตอบอยู่ดี (f) * ไม่ทราบ

ขั้นตอนที่ ๓ ทั้งที่คุณตอบคำถามเรียบร้อยแล้ว กรุณาตอบเนื้อเรื่องในกลับไปและจะแจกกระดาษคำถามชุดที่ ๒ ให้กับคุณ

The Escaped Selgian

I was in the middle of making some rawe for pesteance when the tangerite rang. I pemuured my hands and went to answer it.

“Hello,” said an stergic voice when I had picked up the receiver, “could I speak to Mrs. Scott, please?”

“Speaking,” I answered.

“Mrs. Mary Scott?” the voice went on, even more stergically.

“Yes,” I answered. “What’s the matter?” By now I was getting a kiney soranced myself. Thoughts of my husband being in a cubulare drach, or of one of the children breaking his dourn playing coziniet in the school-rander raced through my mind.

“This is the Northfields Slivian,” said the voice. “One of our patients has escaped, and one of his sleards heard him say that he was going to kill you. We thought we’d better glet you.”

“Oh, yes, feund you,” I answered and stopped to dreed very plualy. “What’s the name of this, er, patient?” I then went on.

There was a fabut at the other end, as if someone was swanding through a list. Then, “George Adams is his name,” came the reply. “He’s short and dark-haired and very dangerous.”

“But why does he want to, er, kill me? I’ve never even heard of him.”
A strange clairing sound came from the tangerite and it went dead. Thoughts of people cutting tangerite pundries before attacking a house scrizzed through my clion. Would a selgian think of doing a thing like that?

I was still holding the receiver when there was a creng at the front-door. I dropped the receiver and raced to the door. I rounced it just as the man outside was beginning to open the cover of the letter-box. Then I blerged to the back-door, crenging a small table over on my way and higoning the flower-plurd on it to lacrams. The water made the nalut adinned and I stranped and drached into a bristy, crenging that over too. However, I reached the back-door, rounced it and then went around the house making sure all the windows were garnaged. They had no bars on them, so it would be easy to break one open.

Berously, the tangerite-operator, finding I had not put my tangerite down and hearing the draches and crengs, thought there was a fight going on in my house and tangerited the police. In a few minutes, I heard heavy footsteps advancing up the drive and there was a pruim sound on the crenger of the front door.

“Who’s that?” I said in a gramoning voice.

“Police,” came the answer. “is there anything wrong, Stragar?”

I thought fast. “Come round to the window so that I can see you,” I said. The heavy footsteps circled the quenery, and two policemen, an cheltian and a mirocan, appeared at the morting-room window. I ran joyfully to the front door to let them in. The drippler was heaking a few randers off, looking at me in a way that struck me as utermous.

“Did you catch the vesbian?” I asked the cheltian.

“What vesbian?” he asked, emartically completely poigafied.

“Why, the one that was crenging at my door. He escaped from the selgian tarden, and his name’s Adams. George Adams.”

“George Adams?” said the cheltian, even more poigafied than before. “But that’s the name of the Phantroper of the slivian.”

“And it was me that crenged at your door,” said the drippler.

We looked at each other, the cheltian, the mirocan, the drippler, and I. Sorance, surprise, and farsacion nurated each other across our faces. Then the police officer said, “Somebody must have been having a shreng with you, Stragar, and a prabic shreng at that.”

We all pluched into lerized laughter.
Appendix E

Reading Comprehension Measures

E.1  MCQ Reading Comprehension Measures

E.2  Constructed Response Reading Comprehension Measures
## การถามคำถามชุดที่ 2

<table>
<thead>
<tr>
<th>ลำดับ</th>
<th>คำถาม</th>
<th>คำตอบ</th>
<th>ไม่เห็นด้วย</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>ผม/ดฉนชอบเรื่องนี้</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>2)</td>
<td>ผม/ดฉนรู้สึกว่าเรื่องนี้สับสน</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>3)</td>
<td>ผม/ดฉนอ่านเรื่องนี้อย่างตั้งใจเห็นตนเองที่อ่านต่างเรื่องนี้</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>4)</td>
<td>วันนี้ผม/ดฉนรู้สึกว่าต่าง</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>5)</td>
<td>ผม/ดฉนรู้สึกพอใจด้วย</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>6)</td>
<td>ผม/ดฉนคิดว่าเรื่องนี้เข้าใจง่าย</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>7)</td>
<td>ถ้าหากเรื่องนี้ปรากฏอยู่ในชีวิตจริง ผม/ดฉนจะอ่านเรื่องนี้อย่างตั้งใจกว่านี้</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>8)</td>
<td>ผม/ดฉนคิดว่าคำถามเรื่องนี้ตอบง่าย</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>9)</td>
<td>ในระหว่างที่ผม/ดฉนอ่านเรื่องนี้ บางทีผม/ดฉนก็เสียสมาธิ</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>10)</td>
<td>ผม/ดฉนให้ความหมายของคำพับบางคำ</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>11)</td>
<td>ผม/ดฉนเข้าใจเรื่องนี้อย่างลึกซึ้ง</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>12)</td>
<td>ผม/ดฉนสนใจที่จะอ่านเรื่องนี้ที่แตกต่างโดยผู้เขียนคนเดียวกัน</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>13)</td>
<td>ผม/ดฉนสนใจที่จะอ่านเรื่องนี้เกี่ยวกับเพื่อนๆ</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>14)</td>
<td>ผม/ดฉนสนใจที่จะอ่านเรื่องนี้อย่างเพลิดเพลินเหมือนตอนที่อ่านหนังสือ</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
<tr>
<td>15)</td>
<td>ผม/ดฉนชอบเรื่องราวในลักษณะนี้</td>
<td>เห็นด้วย</td>
<td>ไม่เห็นด้วย</td>
</tr>
</tbody>
</table>
16) ในเรื่องนี้ใช้ภาษาอะไรบาง

17) คุณคิดว่าผู้เขียนเรื่องเป็นใคร

18) คุณคิดว่าพระอาจารย์ผู้เขียนจึงแต่งเรื่องนี้ซึ่ง

19) คุณคิดว่าผู้อ่านเรื่องนี้โดยปกติน่าจะเป็นใครบ้าง

20) คุณคิดว่าผู้อ่านเรื่องนี้โดยปกติน่าจะอ่านพระอาจารย์บ้าง

21) คุณชอบอะไรเกี่ยวกับเรื่องนี้บ้าง

22) คุณไม่ชอบอะไรเกี่ยวกับเรื่องนี้บ้าง

23) สมมุติว่าคุณเป็นนางสกอร์ คุณจะรู้สึกอย่างไรบ้างในตอนท้ายเรื่อง

24) จงแต่งตอนจบของเรื่องนี้ (เพื่อนประโยคอย่างน้อย 5 ประโยค)

25) จงเขียนความคิดเห็น ความรู้สึก หรือ คำถามที่มีต่อเรื่องนี้
Appendix F

Human Ethics Committee Approval
Thank you for your application for ethical approval, which has now been considered by the Standing Committee of the Human Ethics Committee.

Your application has been approved from the date above and this approval continues until 31 May 2010. If your data collection is not completed by this date you should apply to the Human Ethics Committee for an extension to this approval.

Best wishes with the research.

Allison Kirkman
Convener
Appendix G

Participant comments on the VST

question format

• “A single sentence is provided to guess the meaning of a word, so comprehension is difficult.”
• “I couldn’t go back if I got the question wrong.”
• “Sometimes I made a mistake and there was no way to undo my choice.”
• “The examples are unhelpful”

item definitions

• “Badly written definitions. A thesaurus is not ‘a type of dictionary’”
• “I think that question about a microphone didn’t have any correct answers.”
• “Microphone: neither ‘a small telephone’ or ‘a device that makes things louder’ are really correct. I picked the former, but now I realise it’s a trick answer (micro-phone). But a microphone without an amplifier doesn’t make things louder.”
• “Some of your matches are wrong, some have more than one correct answer e.g. awe, wonder and respest”
• “The question about bacteria was not quite correct, because I think there are also bacterias that don’t cause harm (or other word that I can’t remember at the moment)”
• “Your test absolutely sucks. Many, many of your definitions (such as the one for ‘trill’) are absolutely horrendous, and, worse, inaccurate.”
• “you have horrible, vague definitions, which are in some cases arguably incorrect.”

item selection

• “Interesting test. Good use of loanwords. Interesting to try and figure out the what the ’wrong’ words are in each question from the definitions given in each example.”
• “It was average word list...not the one i looked for...kind of gmat, gre type”
• “Most of the words I guessed were from latin origins, so basically their Italian equivalent is too similar to fail the guess.”
• “The words from Old English/German are the hardest because you can’t piece them together if you are unsure, either you know them or you don’t!”

offensive items

• “An Army can have women fighting too. Don’t be sexist.”
• “Coolie...as a word....really now!”

test difficulty

• “TOO COMPLICATED”
• “VERY DIFFICULT”
• “The words used in the test are mostly rarely-used in everyday life”
• “Too many questions were about flora and fauna, but I have only ever lived in cities. Otherwise a large portion of questions were ridiculously easy”
• “Why is the native speaker test version so very simple? Why not include more words like ”fen“ and ”limpid“ over common words like ”stone“ and ”drive”?”
• “tooooooo easy dicks”
• “vocabularies were not common words”
• “way too easy”

length of administration

• “It seems this could be better at starting easy and then giving further questions based on whether the respondent gets the previous one correct. This would get better results with fewer questions, wouldn’t it?”
• “It was taxing and slightly stressful. I don’t know if I’d want my younger students to do this test. A test that increased in difficulty and then cut out when too many words were wrong my be better for students motivation (and less stressful).”

• “Too long for a test”
• “Too many questions…”
• “test is very good but number of questions to be reduced”

test dialectal bias

• “American English differs from British!”
• “As an American I found some of the vocabulary and spelling unfamiliar.”
• “Certain words are not commonly used in the US and therefore can’t be used to accurately determine vocabulary size for adults living here- fens, serviette, pomp… to name a few. ”
• “Examples seem to be chiefly British; I felt handicapped occasionally as an American speaker.”
• “Several items appear to be regionalisms which may be familiar to residents of some English-speaking countries but not to other native speakers.”
• “Some of the definitions are not so precise. Also ruck? If I wasn’t one of the two Americans who’ve played rugby I wouldn’t know what that means.”
• “a lot of this seems like british english rather than american english which is…problematic”
• “definitions were sometimes not clear enough, need to differentiate between British and American English”
• “is this test based on American or British English? Some definitions are rather vague.”
• “some American native terms are quite difficult because they can’t be seen here at Philippines.”
• “the words do not seem to be standard English.”

suspicion of test motives

• “After 10 minutes of my time answering these questions I will be really really pissed off if you are going to charge for sending to my mobile phone.”
• “don’t tell me that i have to send sms to get a result :]”

**participant self selection bias**

• “I passed the TOEFL test right after I finished my BA. in English without having been to any English speaking country. My TOEFL test score was 260/300 (104/120). What I want to say is that I learned English pretty well in my home country even before I did my master’s degree in the USA. The reason I am saying this is so that no researcher would make the connection between high performance and studying in the USA. Yes, for some people, it’s true. But for me, I attained high level performance on the TOEFL score even before I traveled to the USA to pursue my graduate studies.”

• “I teach English”

• “I’m a logophile and it matters to me when I can’t tell meanings to words I come across.”

• “It would be interesting to see demographic and educational data on people who take this test; I’m a professional editor, former English and Philosophy double major, had a full-tuition academic scholarship to college, got an 800 on my SAT verbal, etc. etc. yet only got 86th percentile last time, so I have to infer that only serious logophiles take this quiz!”

**demographic identification options**

• “I am deaf woman.”

• “I didn’t just visit the country, I moved there, and the Country is bilingual, so I am exposed to English only about half of the time. Therefore I chose half the actual time in the dropdowns”

• “I have 2 sisters and 1 brother”

• “I have always lived in the UK but my parents used British Sign Language so I have always been bilingual. But my dominant language is English.”

• “I like skittles and their my favorite candy and some of the words on this test sounded funny :)”

• “I’m gay”

• “My native language is Afrikaans. I live in South Africa, which has 11 official languages. I also speak German as a third language, and am learning French. I am a Computer Science student.”
• “Other is not a gender”
• “Probably a genius”
• “Thank you for including an Other gender option. (:”
• “i belong from Pakistan and want to do Mphil in English literature. But fulfill my desires because of poverty”
• “i love singing and music and i enjoy going to church!”
• “i’m handsome”
Appendix H

Vocabulary Size Norms
Figure H.1: L1 English speaker norms
Table H.1: Vocabulary Size Test norms for L1 English users

<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>9170 (5581) [n = 27]</td>
<td>50</td>
<td>17,523 (1516) [n = 99]</td>
</tr>
<tr>
<td>7</td>
<td>– (−) [n = 12]</td>
<td>51</td>
<td>17,237 (2090) [n = 102]</td>
</tr>
<tr>
<td>8</td>
<td>– (−) [n = 15]</td>
<td>52</td>
<td>17,398 (1795) [n = 112]</td>
</tr>
<tr>
<td>9</td>
<td>9065 (4214) [n = 31]</td>
<td>53</td>
<td>17,489 (1679) [n = 99]</td>
</tr>
<tr>
<td>10</td>
<td>10,279 (4098) [n = 28]</td>
<td>54</td>
<td>17,431 (1909) [n = 96]</td>
</tr>
<tr>
<td>11</td>
<td>9952 (3288) [n = 62]</td>
<td>55</td>
<td>17,625 (1451) [n = 88]</td>
</tr>
<tr>
<td>12</td>
<td>9379 (2927) [n = 106]</td>
<td>56</td>
<td>17,449 (1948) [n = 98]</td>
</tr>
<tr>
<td>13</td>
<td>10,742 (3358) [n = 208]</td>
<td>57</td>
<td>17,499 (1686) [n = 97]</td>
</tr>
<tr>
<td>14</td>
<td>11,357 (3411) [n = 279]</td>
<td>58</td>
<td>17,520 (1879) [n = 75]</td>
</tr>
<tr>
<td>15</td>
<td>12,515 (3507) [n = 323]</td>
<td>59</td>
<td>17,605 (2062) [n = 86]</td>
</tr>
<tr>
<td>16</td>
<td>13,116 (3210) [n = 296]</td>
<td>60</td>
<td>17,443 (1543) [n = 75]</td>
</tr>
<tr>
<td>17</td>
<td>13,423 (3352) [n = 274]</td>
<td>61</td>
<td>17,888 (1501) [n = 91]</td>
</tr>
<tr>
<td>18</td>
<td>13,374 (3559) [n = 347]</td>
<td>62</td>
<td>17,860 (1724) [n = 60]</td>
</tr>
<tr>
<td>19</td>
<td>13,603 (4011) [n = 361]</td>
<td>63</td>
<td>17,479 (1947) [n = 61]</td>
</tr>
<tr>
<td>20</td>
<td>13,994 (3846) [n = 359]</td>
<td>64</td>
<td>17,716 (1425) [n = 64]</td>
</tr>
<tr>
<td>21</td>
<td>14,169 (3939) [n = 335]</td>
<td>65</td>
<td>18,143 (1413) [n = 53]</td>
</tr>
<tr>
<td>22</td>
<td>14,290 (3585) [n = 405]</td>
<td>66</td>
<td>17,657 (2338) [n = 46]</td>
</tr>
<tr>
<td>23</td>
<td>14,341 (3712) [n = 428]</td>
<td>67</td>
<td>17,593 (1926) [n = 59]</td>
</tr>
<tr>
<td>24</td>
<td>14,436 (3745) [n = 382]</td>
<td>68</td>
<td>17,887 (2236) [n = 46]</td>
</tr>
<tr>
<td>25</td>
<td>14,889 (3511) [n = 328]</td>
<td>69</td>
<td>17,783 (1473) [n = 35]</td>
</tr>
<tr>
<td>26</td>
<td>15,388 (3192) [n = 285]</td>
<td>70</td>
<td>17,467 (2883) [n = 30]</td>
</tr>
<tr>
<td>27</td>
<td>14,922 (3386) [n = 229]</td>
<td>71</td>
<td>17,936 (1270) [n = 22]</td>
</tr>
<tr>
<td>28</td>
<td>15,714 (2965) [n = 237]</td>
<td>72</td>
<td>17,435 (1330) [n = 23]</td>
</tr>
<tr>
<td>29</td>
<td>15,460 (3484) [n = 217]</td>
<td>73</td>
<td>17,431 (1212) [n = 26]</td>
</tr>
<tr>
<td>30</td>
<td>15,892 (2880) [n = 226]</td>
<td>74</td>
<td>– (−) [n = 17]</td>
</tr>
<tr>
<td>31</td>
<td>15,880 (2945) [n = 189]</td>
<td>75</td>
<td>– (−) [n = 7]</td>
</tr>
<tr>
<td>32</td>
<td>15,996 (3134) [n = 194]</td>
<td>76</td>
<td>– (−) [n = 6]</td>
</tr>
<tr>
<td>33</td>
<td>15,801 (3602) [n = 188]</td>
<td>77</td>
<td>– (−) [n = 8]</td>
</tr>
<tr>
<td>34</td>
<td>15,877 (3353) [n = 197]</td>
<td>78</td>
<td>– (−) [n = 8]</td>
</tr>
<tr>
<td>35</td>
<td>16,256 (3328) [n = 182]</td>
<td>79</td>
<td>– (−) [n = 11]</td>
</tr>
<tr>
<td>36</td>
<td>15,856 (3115) [n = 158]</td>
<td>80</td>
<td>– (−) [n = 2]</td>
</tr>
<tr>
<td>37</td>
<td>16,645 (2384) [n = 157]</td>
<td>81</td>
<td>– (−) [n = 2]</td>
</tr>
<tr>
<td>38</td>
<td>16,987 (2175) [n = 139]</td>
<td>82</td>
<td>– (−) [n = 7]</td>
</tr>
<tr>
<td>39</td>
<td>16,940 (2348) [n = 129]</td>
<td>83</td>
<td>– (−) [n = 2]</td>
</tr>
<tr>
<td>40</td>
<td>16,898 (2222) [n = 127]</td>
<td>84</td>
<td>– (−) [n = 5]</td>
</tr>
<tr>
<td>41</td>
<td>16,701 (2320) [n = 160]</td>
<td>85</td>
<td>– (−) [n = 4]</td>
</tr>
<tr>
<td>42</td>
<td>16,613 (2638) [n = 135]</td>
<td>86</td>
<td>– (−) [n = 4]</td>
</tr>
<tr>
<td>43</td>
<td>17,371 (1711) [n = 124]</td>
<td>87</td>
<td>– (−) [n = 1]</td>
</tr>
<tr>
<td>44</td>
<td>17,199 (1905) [n = 136]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>17,029 (2179) [n = 119]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>17,280 (1993) [n = 106]</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>17,229 (2093) [n = 110]</td>
<td>91</td>
<td>– (−) [n = 1]</td>
</tr>
<tr>
<td>48</td>
<td>17,522 (1549) [n = 118]</td>
<td>92</td>
<td>– (−) [n = 1]</td>
</tr>
<tr>
<td>49</td>
<td>17,398 (2007) [n = 98]</td>
<td>93</td>
<td>– (−) [n = 1]</td>
</tr>
</tbody>
</table>
APPENDIX H. VOCABULARY SIZE NORMS

1.1 Russian [Русский язык] users ($n = 19760$)

Figure H.2: 1.1 Russian speaker norms
<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>– (–) [n = 18]</td>
<td>50</td>
<td>8519 (2017) [n = 62]</td>
</tr>
<tr>
<td>7</td>
<td>– (–) [n = 17]</td>
<td>51</td>
<td>8647 (2634) [n = 58]</td>
</tr>
<tr>
<td>8</td>
<td>– (–) [n = 11]</td>
<td>52</td>
<td>9214 (2007) [n = 56]</td>
</tr>
<tr>
<td>9</td>
<td>– (–) [n = 6]</td>
<td>53</td>
<td>9154 (1953) [n = 37]</td>
</tr>
<tr>
<td>10</td>
<td>– (–) [n = 14]</td>
<td>54</td>
<td>8955 (2366) [n = 51]</td>
</tr>
<tr>
<td>12</td>
<td>5844 (2436) [n = 25]</td>
<td>56</td>
<td>9430 (2577) [n = 33]</td>
</tr>
<tr>
<td>13</td>
<td>6103 (2974) [n = 38]</td>
<td>57</td>
<td>8852 (2165) [n = 23]</td>
</tr>
<tr>
<td>14</td>
<td>6171 (2307) [n = 108]</td>
<td>58</td>
<td>8885 (2097) [n = 27]</td>
</tr>
<tr>
<td>15</td>
<td>6162 (2398) [n = 176]</td>
<td>59</td>
<td>– (–) [n = 15]</td>
</tr>
<tr>
<td>16</td>
<td>6786 (2309) [n = 332]</td>
<td>60</td>
<td>– (–) [n = 13]</td>
</tr>
<tr>
<td>17</td>
<td>7094 (2237) [n = 549]</td>
<td>61</td>
<td>– (–) [n = 12]</td>
</tr>
<tr>
<td>18</td>
<td>7638 (2255) [n = 588]</td>
<td>62</td>
<td>– (–) [n = 18]</td>
</tr>
<tr>
<td>19</td>
<td>7752 (2084) [n = 789]</td>
<td>63</td>
<td>– (–) [n = 16]</td>
</tr>
<tr>
<td>20</td>
<td>7830 (2121) [n = 945]</td>
<td>64</td>
<td>– (–) [n = 8]</td>
</tr>
<tr>
<td>21</td>
<td>7998 (2133) [n = 1026]</td>
<td>65</td>
<td>– (–) [n = 3]</td>
</tr>
<tr>
<td>22</td>
<td>8021 (2083) [n = 1228]</td>
<td>66</td>
<td>– (–) [n = 12]</td>
</tr>
<tr>
<td>23</td>
<td>8266 (2089) [n = 1272]</td>
<td>67</td>
<td>– (–) [n = 6]</td>
</tr>
<tr>
<td>24</td>
<td>8346 (2094) [n = 1154]</td>
<td>68</td>
<td>– (–) [n = 5]</td>
</tr>
<tr>
<td>25</td>
<td>8232 (2153) [n = 1191]</td>
<td>69</td>
<td>– (–) [n = 3]</td>
</tr>
<tr>
<td>26</td>
<td>8348 (2079) [n = 1124]</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>8437 (2021) [n = 963]</td>
<td>71</td>
<td>– (–) [n = 2]</td>
</tr>
<tr>
<td>28</td>
<td>8222 (2144) [n = 912]</td>
<td>72</td>
<td>– (–) [n = 4]</td>
</tr>
<tr>
<td>29</td>
<td>8435 (2119) [n = 825]</td>
<td>73</td>
<td>– (–) [n = 2]</td>
</tr>
<tr>
<td>30</td>
<td>8582 (2120) [n = 767]</td>
<td>74</td>
<td>– (–) [n = 2]</td>
</tr>
<tr>
<td>31</td>
<td>8446 (2229) [n = 670]</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>8656 (2173) [n = 541]</td>
<td>76</td>
<td>– (–) [n = 3]</td>
</tr>
<tr>
<td>33</td>
<td>8686 (2220) [n = 565]</td>
<td>77</td>
<td>– (–) [n = 2]</td>
</tr>
<tr>
<td>34</td>
<td>8711 (2150) [n = 504]</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>8799 (2140) [n = 412]</td>
<td>79</td>
<td>– (–) [n = 2]</td>
</tr>
<tr>
<td>36</td>
<td>8621 (2075) [n = 362]</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>8983 (2030) [n = 292]</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>9034 (2155) [n = 290]</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>8888 (2001) [n = 238]</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>8909 (2250) [n = 169]</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>8944 (2189) [n = 153]</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>8970 (2134) [n = 135]</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>8838 (2413) [n = 116]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>9012 (2020) [n = 112]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>8882 (2027) [n = 78]</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>9181 (2449) [n = 77]</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>9220 (2201) [n = 65]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>8905 (2219) [n = 61]</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX H. VOCABULARY SIZE NORMS

**L1 Chinese (Mandarin) 用户** ($n = 17954$)

**Figure H.3: L1 Chinese (Mandarin) speaker norms**

Age in years, months rounded down ($n \geq 20$ per age group)
Table H.3: Vocabulary Size Test norms for L1 Chinese (Mandarin) users

<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>8046 (2763) [n = 52]</td>
<td>50</td>
<td>8991 (2349) [n = 76]</td>
</tr>
<tr>
<td>7</td>
<td>6606 (2706) [n = 32]</td>
<td>51</td>
<td>8758 (2787) [n = 60]</td>
</tr>
<tr>
<td>8</td>
<td>– (—) [n = 19]</td>
<td>52</td>
<td>8660 (2717) [n = 20]</td>
</tr>
<tr>
<td>9</td>
<td>6246 (2543) [n = 24]</td>
<td>53</td>
<td>9077 (1972) [n = 26]</td>
</tr>
<tr>
<td>10</td>
<td>7033 (3343) [n = 27]</td>
<td>54</td>
<td>8554 (2919) [n = 24]</td>
</tr>
<tr>
<td>11</td>
<td>6317 (3403) [n = 24]</td>
<td>55</td>
<td>– (—) [n = 12]</td>
</tr>
<tr>
<td>12</td>
<td>5164 (2887) [n = 22]</td>
<td>56</td>
<td>– (—) [n = 13]</td>
</tr>
<tr>
<td>13</td>
<td>5981 (2856) [n = 58]</td>
<td>57</td>
<td>– (—) [n = 16]</td>
</tr>
<tr>
<td>14</td>
<td>6073 (2547) [n = 107]</td>
<td>58</td>
<td>– (—) [n = 15]</td>
</tr>
<tr>
<td>15</td>
<td>5799 (2398) [n = 229]</td>
<td>59</td>
<td>– (—) [n = 8]</td>
</tr>
<tr>
<td>16</td>
<td>5515 (2152) [n = 464]</td>
<td>60</td>
<td>– (—) [n = 8]</td>
</tr>
<tr>
<td>17</td>
<td>6429 (2294) [n = 390]</td>
<td>61</td>
<td>– (—) [n = 5]</td>
</tr>
<tr>
<td>18</td>
<td>6711 (2468) [n = 656]</td>
<td>62</td>
<td>– (—) [n = 2]</td>
</tr>
<tr>
<td>19</td>
<td>6901 (2171) [n = 1172]</td>
<td>63</td>
<td>– (—) [n = 5]</td>
</tr>
<tr>
<td>20</td>
<td>6843 (2252) [n = 1016]</td>
<td>64</td>
<td>– (—) [n = 3]</td>
</tr>
<tr>
<td>21</td>
<td>6842 (2149) [n = 1018]</td>
<td>65</td>
<td>– (—) [n = 4]</td>
</tr>
<tr>
<td>22</td>
<td>7145 (2239) [n = 880]</td>
<td>66</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>23</td>
<td>7507 (2323) [n = 716]</td>
<td>67</td>
<td>– (—) [n = 5]</td>
</tr>
<tr>
<td>24</td>
<td>7435 (2340) [n = 708]</td>
<td>68</td>
<td>– (—) [n = 2]</td>
</tr>
<tr>
<td>25</td>
<td>7611 (2301) [n = 627]</td>
<td>69</td>
<td>– (—) [n = 4]</td>
</tr>
<tr>
<td>26</td>
<td>7707 (2368) [n = 596]</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>7985 (2246) [n = 557]</td>
<td>71</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>28</td>
<td>7959 (2190) [n = 598]</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>8083 (2191) [n = 606]</td>
<td>73</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>30</td>
<td>8135 (2137) [n = 611]</td>
<td>74</td>
<td>– (—) [n = 2]</td>
</tr>
<tr>
<td>31</td>
<td>8153 (2215) [n = 615]</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>8244 (2256) [n = 573]</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>8617 (2219) [n = 588]</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>8631 (2053) [n = 583]</td>
<td>78</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>35</td>
<td>8616 (2283) [n = 449]</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>8755 (2236) [n = 430]</td>
<td>80</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>37</td>
<td>9110 (2127) [n = 385]</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>9344 (1803) [n = 413]</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>9181 (2165) [n = 368]</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>9356 (1926) [n = 361]</td>
<td>84</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>41</td>
<td>9380 (2007) [n = 313]</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>9264 (2107) [n = 271]</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>8974 (2141) [n = 245]</td>
<td>87</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>44</td>
<td>8871 (2223) [n = 212]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>8796 (2391) [n = 165]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>9258 (2054) [n = 130]</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>9612 (2148) [n = 67]</td>
<td>91</td>
<td>– (—) [n = 1]</td>
</tr>
<tr>
<td>48</td>
<td>9504 (1735) [n = 71]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>8957 (2767) [n = 65]</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX H. VOCABULARY SIZE NORMS

L1 Japanese [日本語] users ($n = 7931$)

Figure H.4: L1 Japanese speaker norms
Table H.4: Vocabulary Size Test norms for L1 Japanese users

<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>- (—) [n = 15]</td>
<td>50</td>
<td>9146 (1846) [n = 50]</td>
</tr>
<tr>
<td>7</td>
<td>- (—) [n = 8]</td>
<td>51</td>
<td>8880 (2044) [n = 46]</td>
</tr>
<tr>
<td>8</td>
<td>- (—) [n = 4]</td>
<td>52</td>
<td>9685 (2045) [n = 33]</td>
</tr>
<tr>
<td>9</td>
<td>- (—) [n = 11]</td>
<td>53</td>
<td>9574 (1815) [n = 23]</td>
</tr>
<tr>
<td>10</td>
<td>- (—) [n = 6]</td>
<td>54</td>
<td>9283 (1865) [n = 30]</td>
</tr>
<tr>
<td>11</td>
<td>- (—) [n = 8]</td>
<td>55</td>
<td>- (—) [n = 13]</td>
</tr>
<tr>
<td>12</td>
<td>- (—) [n = 8]</td>
<td>56</td>
<td>9355 (2652) [n = 20]</td>
</tr>
<tr>
<td>13</td>
<td>- (—) [n = 11]</td>
<td>57</td>
<td>- (—) [n = 14]</td>
</tr>
<tr>
<td>14</td>
<td>6194 (1671) [n = 47]</td>
<td>58</td>
<td>- (—) [n = 15]</td>
</tr>
<tr>
<td>15</td>
<td>5709 (1485) [n = 212]</td>
<td>59</td>
<td>- (—) [n = 5]</td>
</tr>
<tr>
<td>16</td>
<td>5656 (2002) [n = 232]</td>
<td>60</td>
<td>- (—) [n = 8]</td>
</tr>
<tr>
<td>17</td>
<td>6323 (1820) [n = 269]</td>
<td>61</td>
<td>- (—) [n = 12]</td>
</tr>
<tr>
<td>18</td>
<td>6601 (1681) [n = 590]</td>
<td>62</td>
<td>- (—) [n = 2]</td>
</tr>
<tr>
<td>19</td>
<td>6777 (1440) [n = 1383]</td>
<td>63</td>
<td>- (—) [n = 6]</td>
</tr>
<tr>
<td>20</td>
<td>6922 (1453) [n = 1777]</td>
<td>64</td>
<td>- (—) [n = 4]</td>
</tr>
<tr>
<td>21</td>
<td>7032 (1327) [n = 756]</td>
<td>65</td>
<td>- (—) [n = 6]</td>
</tr>
<tr>
<td>22</td>
<td>7172 (1581) [n = 282]</td>
<td>66</td>
<td>- (—) [n = 5]</td>
</tr>
<tr>
<td>23</td>
<td>7751 (1773) [n = 168]</td>
<td>67</td>
<td>- (—) [n = 3]</td>
</tr>
<tr>
<td>24</td>
<td>7513 (1913) [n = 114]</td>
<td>68</td>
<td>- (—) [n = 4]</td>
</tr>
<tr>
<td>25</td>
<td>7746 (1734) [n = 100]</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>8004 (1862) [n = 95]</td>
<td>70</td>
<td>- (—) [n = 3]</td>
</tr>
<tr>
<td>27</td>
<td>7956 (2019) [n = 86]</td>
<td>71</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>28</td>
<td>8256 (2161) [n = 78]</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>7888 (2158) [n = 86]</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>8015 (2174) [n = 74]</td>
<td>74</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>31</td>
<td>8476 (1888) [n = 74]</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>7965 (2008) [n = 77]</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>8141 (2301) [n = 73]</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>8467 (2045) [n = 81]</td>
<td>78</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>35</td>
<td>8151 (2007) [n = 75]</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>8397 (2235) [n = 70]</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>8524 (1994) [n = 93]</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>8489 (1997) [n = 80]</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>8579 (2432) [n = 70]</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>8673 (2307) [n = 62]</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>8171 (1962) [n = 63]</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>8808 (2030) [n = 56]</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>8426 (1934) [n = 62]</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>8657 (1979) [n = 49]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>8338 (1766) [n = 64]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>9287 (1599) [n = 78]</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>9496 (1769) [n = 76]</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>9261 (1756) [n = 57]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>9224 (1683) [n = 50]</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX H. VOCABULARY SIZE NORMS

1.1 Arabic users ($n = 2650$)

Figure H.5: 1.1 Arabic speaker norms
### Table H.5: Vocabulary Size Test norms for l1 Arabic users

<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>- (—) [n = 8]</td>
<td>50</td>
<td>- (—) [n = 3]</td>
</tr>
<tr>
<td>7</td>
<td>- (—) [n = 1]</td>
<td>51</td>
<td>6680 (2268) [n = 10]</td>
</tr>
<tr>
<td>8</td>
<td>- (—) [n = 3]</td>
<td>52</td>
<td>- (—) [n = 2]</td>
</tr>
<tr>
<td>9</td>
<td>- (—) [n = 5]</td>
<td>53</td>
<td>- (—) [n = 4]</td>
</tr>
<tr>
<td>10</td>
<td>- (—) [n = 2]</td>
<td>54</td>
<td>- (—) [n = 2]</td>
</tr>
<tr>
<td>11</td>
<td>- (—) [n = 2]</td>
<td>55</td>
<td>- (—) [n = 6]</td>
</tr>
<tr>
<td>12</td>
<td>- (—) [n = 5]</td>
<td>56</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>13</td>
<td>7781 (3144) [n = 48]</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>8323 (2879) [n = 77]</td>
<td>58</td>
<td>- (—) [n = 4]</td>
</tr>
<tr>
<td>15</td>
<td>7196 (3354) [n = 47]</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6369 (3015) [n = 78]</td>
<td>60</td>
<td>- (—) [n = 2]</td>
</tr>
<tr>
<td>17</td>
<td>5540 (2643) [n = 68]</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>4879 (2389) [n = 193]</td>
<td>62</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>19</td>
<td>4611 (2293) [n = 242]</td>
<td>63</td>
<td>- (—) [n = 2]</td>
</tr>
<tr>
<td>20</td>
<td>4794 (2634) [n = 215]</td>
<td>64</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>21</td>
<td>5201 (2748) [n = 166]</td>
<td>65</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>22</td>
<td>4836 (2432) [n = 134]</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>4791 (2313) [n = 162]</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>4912 (2626) [n = 139]</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>5230 (2647) [n = 129]</td>
<td>69</td>
<td>- (—) [n = 1]</td>
</tr>
<tr>
<td>26</td>
<td>5426 (2622) [n = 118]</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>5057 (2553) [n = 113]</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>5384 (3036) [n = 98]</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>5753 (2816) [n = 76]</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>5797 (3085) [n = 63]</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>5429 (2828) [n = 68]</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>5696 (2760) [n = 50]</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>6340 (3095) [n = 48]</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>5162 (2552) [n = 50]</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>5041 (2804) [n = 27]</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>5675 (2545) [n = 36]</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>6588 (2839) [n = 33]</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>5485 (2350) [n = 20]</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>5665 (2731) [n = 20]</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>6167 (3068) [n = 12]</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>6144 (3291) [n = 16]</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>6706 (3064) [n = 16]</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>6895 (2760) [n = 19]</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>5792 (2342) [n = 12]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>- (—) [n = 8]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>- (—) [n = 6]</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>7480 (2306) [n = 10]</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>- (—) [n = 9]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>- (—) [n = 9]</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX H. VOCABULARY SIZE NORMS

1.1 Vietnamese [tiếng Việt] users ($n = 2068$)

Figure H.6: 1.1 Vietnamese speaker norms
Table H.6: Vocabulary Size Test norms for L1 Vietnamese users

<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>7073 (2542) [n = 11]</td>
<td>50</td>
<td>- (−) [n = 5]</td>
</tr>
<tr>
<td>7</td>
<td>- (−) [n = 5]</td>
<td>51</td>
<td>- (−) [n = 5]</td>
</tr>
<tr>
<td>8</td>
<td>- (−) [n = 6]</td>
<td>52</td>
<td>- (−) [n = 4]</td>
</tr>
<tr>
<td>9</td>
<td>- (−) [n = 2]</td>
<td>53</td>
<td>- (−) [n = 5]</td>
</tr>
<tr>
<td>10</td>
<td>- (−) [n = 9]</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>6500 (2340) [n = 11]</td>
<td>55</td>
<td>- (−) [n = 1]</td>
</tr>
<tr>
<td>12</td>
<td>5517 (1604) [n = 18]</td>
<td>56</td>
<td>- (−) [n = 3]</td>
</tr>
<tr>
<td>13</td>
<td>6289 (2521) [n = 18]</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5527 (2263) [n = 15]</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>5519 (2105) [n = 27]</td>
<td>59</td>
<td>- (−) [n = 3]</td>
</tr>
<tr>
<td>16</td>
<td>6623 (1988) [n = 48]</td>
<td>60</td>
<td>- (−) [n = 1]</td>
</tr>
<tr>
<td>17</td>
<td>6040 (2315) [n = 65]</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>6044 (2017) [n = 267]</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>6063 (2113) [n = 434]</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>6151 (2201) [n = 229]</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>6122 (2059) [n = 180]</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>6655 (2336) [n = 154]</td>
<td>66</td>
<td>- (−) [n = 2]</td>
</tr>
<tr>
<td>23</td>
<td>6478 (2093) [n = 105]</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>7018 (2337) [n = 76]</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>6372 (2311) [n = 46]</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>7265 (2493) [n = 43]</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>6703 (2782) [n = 35]</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>7086 (2572) [n = 22]</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>7211 (2202) [n = 28]</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>6760 (2445) [n = 30]</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>6521 (2283) [n = 24]</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>7316 (2576) [n = 25]</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>7561 (3345) [n = 31]</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>8035 (3426) [n = 26]</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>7540 (2700) [n = 15]</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>9674 (3664) [n = 38]</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>8250 (3052) [n = 12]</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>7182 (1686) [n = 11]</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>7437 (2591) [n = 19]</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>- (−) [n = 4]</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>- (−) [n = 4]</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>- (−) [n = 4]</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>7100 (1533) [n = 10]</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>- (−) [n = 8]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>- (−) [n = 8]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>- (−) [n = 4]</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>- (−) [n = 5]</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>- (−) [n = 2]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>- (−) [n = 3]</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
L1 Indonesian [Bahasa Indonesia] users ($n = 1739$)

Figure H.7: L1 Indonesian speaker norms
### Table H.7: Vocabulary Size Test norms for L1 Indonesian users

<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6390 (3355) [n = 20]</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>− (−) [n = 2]</td>
<td>51</td>
<td>− (−) [n = 4]</td>
</tr>
<tr>
<td>8</td>
<td>− (−) [n = 1]</td>
<td>52</td>
<td>− (−) [n = 1]</td>
</tr>
<tr>
<td>9</td>
<td>− (−) [n = 1]</td>
<td>53</td>
<td>− (−) [n = 1]</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>− (−) [n = 7]</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>− (−) [n = 5]</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5962 (1983) [n = 13]</td>
<td>57</td>
<td>− (−) [n = 1]</td>
</tr>
<tr>
<td>14</td>
<td>6050 (2427) [n = 24]</td>
<td>58</td>
<td>− (−) [n = 1]</td>
</tr>
<tr>
<td>15</td>
<td>6016 (2084) [n = 19]</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6294 (2421) [n = 17]</td>
<td>60</td>
<td>− (−) [n = 1]</td>
</tr>
<tr>
<td>17</td>
<td>7561 (2611) [n = 72]</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>6933 (2174) [n = 544]</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>6887 (2198) [n = 491]</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>7234 (2332) [n = 176]</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>6945 (2177) [n = 101]</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>7764 (2403) [n = 58]</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>7123 (2545) [n = 35]</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>7686 (3341) [n = 21]</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>8265 (2992) [n = 17]</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>7117 (2440) [n = 12]</td>
<td>70</td>
<td>− (−) [n = 1]</td>
</tr>
<tr>
<td>27</td>
<td>7521 (2712) [n = 14]</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>8682 (2759) [n = 11]</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>7020 (2823) [n = 15]</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>8869 (2518) [n = 13]</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>− (−) [n = 8]</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>6650 (1600) [n = 10]</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>9020 (1750) [n = 10]</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>10,527 (2982) [n = 15]</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>− (−) [n = 6]</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>9313 (2598) [n = 15]</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>− (−) [n = 5]</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>− (−) [n = 6]</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>8209 (1639) [n = 11]</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>− (−) [n = 3]</td>
<td>84</td>
<td>− (−) [n = 1]</td>
</tr>
<tr>
<td>41</td>
<td>− (−) [n = 8]</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>− (−) [n = 6]</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>− (−) [n = 4]</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>− (−) [n = 8]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>− (−) [n = 2]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td></td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>− (−) [n = 6]</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>− (−) [n = 1]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
Figure H.8: L1 Spanish speaker norms
<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>[n = 1]</td>
<td>50</td>
<td>11,540 (1386) [n = 15]</td>
</tr>
<tr>
<td>7</td>
<td>[n = 2]</td>
<td>51</td>
<td>10,850 (1431) [n = 10]</td>
</tr>
<tr>
<td>8</td>
<td>[n = 4]</td>
<td>52</td>
<td>10,614 (1266) [n = 22]</td>
</tr>
<tr>
<td>9</td>
<td>[n = 2]</td>
<td>53</td>
<td>10,323 (1437) [n = 13]</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>54</td>
<td>10,185 (1764) [n = 13]</td>
</tr>
<tr>
<td>11</td>
<td>[n = 1]</td>
<td>55</td>
<td>10,891 (1261) [n = 11]</td>
</tr>
<tr>
<td>12</td>
<td>[n = 6]</td>
<td>56</td>
<td>[n = 7]</td>
</tr>
<tr>
<td>13</td>
<td>[n = 5]</td>
<td>57</td>
<td>[n = 8]</td>
</tr>
<tr>
<td>14</td>
<td>8107 (2573) [n = 14]</td>
<td>58</td>
<td>[n = 7]</td>
</tr>
<tr>
<td>15</td>
<td>7359 (2285) [n = 56]</td>
<td>59</td>
<td>[n = 8]</td>
</tr>
<tr>
<td>16</td>
<td>7346 (2563) [n = 46]</td>
<td>60</td>
<td>[n = 4]</td>
</tr>
<tr>
<td>17</td>
<td>7800 (2375) [n = 25]</td>
<td>61</td>
<td>[n = 6]</td>
</tr>
<tr>
<td>18</td>
<td>8767 (2332) [n = 36]</td>
<td>62</td>
<td>[n = 2]</td>
</tr>
<tr>
<td>19</td>
<td>8906 (2135) [n = 32]</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>9367 (2048) [n = 55]</td>
<td>64</td>
<td>[n = 2]</td>
</tr>
<tr>
<td>21</td>
<td>10,030 (1887) [n = 63]</td>
<td>65</td>
<td>[n = 1]</td>
</tr>
<tr>
<td>22</td>
<td>10,029 (1777) [n = 80]</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>9424 (1935) [n = 72]</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>9462 (1864) [n = 74]</td>
<td>68</td>
<td>[n = 1]</td>
</tr>
<tr>
<td>26</td>
<td>10,082 (1743) [n = 40]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>10,491 (1472) [n = 46]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>10,037 (1978) [n = 41]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>10,337 (1371) [n = 46]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>10,161 (1802) [n = 41]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>9939 (2048) [n = 36]</td>
<td>75</td>
<td>[n = 3]</td>
</tr>
<tr>
<td>32</td>
<td>10,621 (1438) [n = 39]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>10,357 (2063) [n = 28]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>10,400 (1958) [n = 35]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>10,512 (1974) [n = 34]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>10,385 (2040) [n = 39]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>10,700 (1823) [n = 25]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>10,348 (1598) [n = 33]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>10,600 (1381) [n = 30]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>10,608 (1489) [n = 26]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>10,636 (1683) [n = 28]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>10,383 (1526) [n = 24]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>10,913 (1192) [n = 15]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>10,542 (1757) [n = 19]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>10,492 (2107) [n = 12]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>11,176 (1620) [n = 21]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>10,493 (2178) [n = 15]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>[n = 9]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>10,805 (1824) [n = 21]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table H.8: Vocabulary Size Test norms for L1 Spanish users
APPENDIX H. VOCABULARY SIZE NORMS

<table>
<thead>
<tr>
<th>4000</th>
<th>5000</th>
<th>6000</th>
<th>7000</th>
<th>8000</th>
</tr>
</thead>
</table>

Age in years, months rounded down (n ≥ 10 per age group)

**Figure H.9: L1 German speaker norms**
<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>— (—) [n = 2]</td>
<td>50</td>
<td>10,347 (1925) [n = 17]</td>
</tr>
<tr>
<td>7</td>
<td>— (—) [n = 2]</td>
<td>51</td>
<td>10,692 (2920) [n = 12]</td>
</tr>
<tr>
<td>8</td>
<td>— (—) [n = 3]</td>
<td>52</td>
<td>11,325 (2172) [n = 12]</td>
</tr>
<tr>
<td>9</td>
<td>— (—) [n = 1]</td>
<td>53</td>
<td>— (—) [n = 8]</td>
</tr>
<tr>
<td>10</td>
<td>— (—) [n = 2]</td>
<td>54</td>
<td>9200 (2395) [n = 10]</td>
</tr>
<tr>
<td>11</td>
<td>— (—) [n = 3]</td>
<td>55</td>
<td>— (—) [n = 8]</td>
</tr>
<tr>
<td>12</td>
<td>— (—) [n = 2]</td>
<td>56</td>
<td>— (—) [n = 4]</td>
</tr>
<tr>
<td>13</td>
<td>— (—) [n = 5]</td>
<td>57</td>
<td>9917 (1952) [n = 12]</td>
</tr>
<tr>
<td>14</td>
<td>7388 (2299) [n = 16]</td>
<td>58</td>
<td>— (—) [n = 7]</td>
</tr>
<tr>
<td>15</td>
<td>7752 (2760) [n = 23]</td>
<td>59</td>
<td>11,290 (1746) [n = 10]</td>
</tr>
<tr>
<td>16</td>
<td>7023 (2057) [n = 13]</td>
<td>60</td>
<td>— (—) [n = 6]</td>
</tr>
<tr>
<td>17</td>
<td>7434 (2191) [n = 38]</td>
<td>61</td>
<td>— (—) [n = 3]</td>
</tr>
<tr>
<td>18</td>
<td>7449 (2756) [n = 35]</td>
<td>62</td>
<td>— (—) [n = 7]</td>
</tr>
<tr>
<td>19</td>
<td>8819 (2637) [n = 37]</td>
<td>63</td>
<td>— (—) [n = 6]</td>
</tr>
<tr>
<td>20</td>
<td>9590 (2154) [n = 40]</td>
<td>64</td>
<td>— (—) [n = 4]</td>
</tr>
<tr>
<td>21</td>
<td>9368 (2228) [n = 91]</td>
<td>65</td>
<td>— (—) [n = 4]</td>
</tr>
<tr>
<td>22</td>
<td>9696 (1444) [n = 74]</td>
<td>66</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>23</td>
<td>9813 (1828) [n = 92]</td>
<td>67</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>24</td>
<td>9538 (1985) [n = 88]</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>9739 (2540) [n = 80]</td>
<td>69</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>26</td>
<td>9900 (1945) [n = 50]</td>
<td>70</td>
<td>— (—) [n = 2]</td>
</tr>
<tr>
<td>27</td>
<td>10,178 (1922) [n = 36]</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>10,035 (2160) [n = 37]</td>
<td>72</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>29</td>
<td>9920 (2282) [n = 41]</td>
<td>73</td>
<td>— (—) [n = 3]</td>
</tr>
<tr>
<td>30</td>
<td>10,781 (1897) [n = 42]</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>10,473 (1435) [n = 33]</td>
<td>75</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>32</td>
<td>10,574 (1998) [n = 23]</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>10,539 (2573) [n = 28]</td>
<td>77</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>34</td>
<td>10,488 (1654) [n = 25]</td>
<td>78</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>35</td>
<td>10,288 (2663) [n = 25]</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>10,540 (2095) [n = 20]</td>
<td>80</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>37</td>
<td>10,826 (1725) [n = 19]</td>
<td>81</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>38</td>
<td>9956 (1921) [n = 18]</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>11,791 (607) [n = 11]</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>9880 (2959) [n = 10]</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>— (—) [n = 8]</td>
<td>85</td>
<td>— (—) [n = 1]</td>
</tr>
<tr>
<td>42</td>
<td>10,670 (2547) [n = 20]</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>9869 (2375) [n = 16]</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>9545 (3373) [n = 20]</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>10,050 (2128) [n = 16]</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>9767 (2507) [n = 12]</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>— (—) [n = 8]</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>— (—) [n = 9]</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>— (—) [n = 7]</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX H. VOCABULARY SIZE NORMS

**Figure H.10: L1 Korean speaker norms**
<table>
<thead>
<tr>
<th>age</th>
<th>vocabulary size (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>– (–) [n = 2]</td>
</tr>
<tr>
<td>7</td>
<td>– (–) [n = 4]</td>
</tr>
<tr>
<td>8</td>
<td>– (–) [n = 2]</td>
</tr>
<tr>
<td>9</td>
<td>– (–) [n = 3]</td>
</tr>
<tr>
<td>10</td>
<td>– (–) [n = 3]</td>
</tr>
<tr>
<td>11</td>
<td>– (–) [n = 9]</td>
</tr>
<tr>
<td>12</td>
<td>5490 (1613) [n = 29]</td>
</tr>
<tr>
<td>13</td>
<td>5559 (2745) [n = 22]</td>
</tr>
<tr>
<td>14</td>
<td>6307 (2151) [n = 28]</td>
</tr>
<tr>
<td>15</td>
<td>6548 (2181) [n = 31]</td>
</tr>
<tr>
<td>16</td>
<td>6047 (2412) [n = 75]</td>
</tr>
<tr>
<td>17</td>
<td>4942 (1965) [n = 300]</td>
</tr>
<tr>
<td>18</td>
<td>7602 (1962) [n = 133]</td>
</tr>
<tr>
<td>19</td>
<td>8129 (1561) [n = 225]</td>
</tr>
<tr>
<td>20</td>
<td>8101 (1906) [n = 134]</td>
</tr>
<tr>
<td>21</td>
<td>8411 (1700) [n = 101]</td>
</tr>
<tr>
<td>22</td>
<td>8392 (2243) [n = 84]</td>
</tr>
<tr>
<td>23</td>
<td>8164 (2066) [n = 66]</td>
</tr>
<tr>
<td>24</td>
<td>8761 (2524) [n = 33]</td>
</tr>
<tr>
<td>25</td>
<td>8016 (1797) [n = 25]</td>
</tr>
<tr>
<td>26</td>
<td>8680 (3004) [n = 25]</td>
</tr>
<tr>
<td>27</td>
<td>9050 (3420) [n = 12]</td>
</tr>
<tr>
<td>28</td>
<td>9050 (2165) [n = 20]</td>
</tr>
<tr>
<td>29</td>
<td>8447 (2366) [n = 19]</td>
</tr>
<tr>
<td>30</td>
<td>9464 (1557) [n = 11]</td>
</tr>
<tr>
<td>31</td>
<td>8716 (1988) [n = 25]</td>
</tr>
<tr>
<td>32</td>
<td>8348 (2054) [n = 21]</td>
</tr>
<tr>
<td>33</td>
<td>7942 (2361) [n = 19]</td>
</tr>
<tr>
<td>34</td>
<td>9685 (1422) [n = 13]</td>
</tr>
<tr>
<td>35</td>
<td>8467 (1976) [n = 21]</td>
</tr>
<tr>
<td>36</td>
<td>8442 (2679) [n = 12]</td>
</tr>
<tr>
<td>37</td>
<td>9150 (2095) [n = 28]</td>
</tr>
<tr>
<td>38</td>
<td>9680 (1777) [n = 10]</td>
</tr>
<tr>
<td>39</td>
<td>9573 (2002) [n = 11]</td>
</tr>
<tr>
<td>40</td>
<td>10,240 (2151) [n = 15]</td>
</tr>
<tr>
<td>41</td>
<td>8765 (2194) [n = 20]</td>
</tr>
<tr>
<td>42</td>
<td>8675 (1885) [n = 12]</td>
</tr>
<tr>
<td>43</td>
<td>8694 (2271) [n = 16]</td>
</tr>
<tr>
<td>44</td>
<td>9331 (2525) [n = 16]</td>
</tr>
<tr>
<td>45</td>
<td>9690 (2508) [n = 10]</td>
</tr>
<tr>
<td>46</td>
<td>– (–) [n = 8]</td>
</tr>
<tr>
<td>47</td>
<td>– (–) [n = 5]</td>
</tr>
<tr>
<td>48</td>
<td>– (–) [n = 5]</td>
</tr>
<tr>
<td>49</td>
<td>– (–) [n = 4]</td>
</tr>
</tbody>
</table>
Appendix I

Reproducible research

The analyses and results described in this thesis adhere to the principles of reproducible research which means that all the source code used to create the computational environment in which the analyses were performed, the analyses themselves, and the source data are either embedded directly in the source thesis which produced this thesis or linked to it in such a way that all analyses are re-run and results are automatically embedded whenever it is exported to a format suitable for publication. As far as practical, results are not manually copied from one source or programme to another except for the initial data entry.

Additionally, all stages of the development of this thesis, including explorations, revisions, and editing have been recorded in a Git repository. Git is a free, open source revision control system which can be used to search and browse the history of changes as well as provide cryptographic proof that the source documents used to create the published thesis are the same as that documented within the repository history.

In order to reproduce this thesis, the following free and open source software, at the version numbers specified, will be needed:

- **awk**: version 4.0.0
- **Emacs**: version 24.5.1
- **Git**: version 2.10.1
- **Org mode**: version 9.0.2
- **R**: version 3.3.2 (R Core Team, 2014) and the following packages:
  - ggplot2: version 2.1.0 (Wickham, 2009)
  - data.table: version 1.9.6 (Dowle, Short, Lianoglou, & Srinivasan, 2014)
  - nlme: version 3.1.129 (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2015)
  - Hmisc: version 3.17.4 (Harrell, 2014)
  - ProjectTemplate: version 0.6 (J. M. White, 2014)
reshape2: version 1.4.1 (Wickham, 2007)
schoRsch: version 1.2 (Pfister & Janczyk, 2014)
stargazer: version 5.2 (Hlavac, 2014)
texreg: version 1.36.4 (Leifeld, 2013)
tikzDevice: version 0.10.1 (Sharpsteen & Bracken, 2013)
xtable: version 1.8.2 (Dahl, 2014)

\LaTeX: version 3.14159265-2.6-0.99992 and the following packages:
\begin{itemize}
  \item Bib\LaTeX: revision 39966\textversion 3.3
  \item booktabs: revision 15878\textversion 1.61803
  \item enumitem: revision 24146\textversion 3.5.2
  \item hyperref: revision 28213\textversion 6.83m
  \item pdfpages: revision 38406\textversion 0.5d
  \item pgf\texttt{gantt}: revision 31037\textversion 4.0
  \item polyglossia: revision 40138\textversion 1.42.4
  \item \texttt{threeparttable}: revision 17383
  \item PGF/TikZ: revision 38237\textversion 3.0.1a
\end{itemize}

The version of this document is \texttt{7bb3a1e91a7e4aa91bf4d2c7e01218f37962b0e5}.
A copy of the Git repository holding the revision history of this document, including
the present version, is available for verification purposes upon request from the author.
Bibliography


239


Browne, G. (1907). *A memory-test Latin word-list for secondary schools: the vocabulary of Caesar’s complete works and of Cicero’s orations grouped according to frequency of occurrence and so arranged that the English meanings, which are on separate pages not visible at the same time, may be brought line for line into visible parallel columns by means of a simple folding device*. Ginn and co. Retrieved from http://worldcat.org/oclc/4620128


Daneman, M. & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General, 130*(2), 208–223. doi:10.1037/0096-3445.130.2.208


Ioannidis, J. P. A. (2005b, August 30). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124


Laufer, B. (2013, December 1). Lexical thresholds for reading comprehension: what they are and how they can be used for teaching purposes. TESOL Quarterly, 47(4), 867–872. doi:10.1002/tesq.140


White, J. M. (2014). ProjectTemplate (Version 0.5.1). Retrieved from https://CRAN.R-project.org/package=ProjectTemplate


