

ABSTRACT

Title of Dissertation: TOWARD A THEORY-BASED ACCOUNT
OF THE L2 VOCABULARY PROCESSING
AND LEARNING BENEFITS OF READING
WHILE LISTENING

Jonathan Edward Malone, Doctor of Philosophy,
2024

Dissertation directed by: Professor Kira Gor, Second Language
Acquisition (Co-Chair)

Assistant Professor Bronson Hui, Second
Language Acquisition (Co-Chair)

The tantalizing prospects of learning benefits from multimodal conditions on second language (L2) learning in general, and L2 vocabulary development in particular, have important implications. Indeed, opening a language learning app on any device provides the immediate experience of simultaneous input modalities, and a wide range of input types. But how helpful is multimodality to vocabulary learning, especially when the focus of the learner is on the meaning of a text? Researchers have manipulated input to compare a variety of learning conditions and examined vocabulary learning gains. However, relatively few within second language acquisition (SLA) have utilized real-time monitoring of learner behavior to examine how learners encounter new words over multiple exposures during a reading task, and how the quality of these encounters may or may not influence explicit learning outcomes. Even fewer have mapped differences in the developmental trajectory of form-form and form-meaning mapping for new words at the group level, comparing reading only (RO) with reading while listening (RWL). Crucially, to my knowledge, none have made or tested predictions within RWL on possible

psycholinguistic source(s) of reported benefits. Our understanding of outcome benefits, along with implications for optimizing input in classroom or individual instructed contexts, is thereby quite limited.

My dissertation study was designed to address each of these issues. 119 advanced English learners read or read while listening to a 7,400-word short story under incidental conditions (time pressure, focus on comprehension, and unannounced posttest outcomes). The text was embedded with 25 target pseudoword items 10 times each, with target items replacing real nouns in object positions. Measures of real-time form learning were defined as faster reading times and fewer total visits to the new words across encounters (Godfroid, 2020b), and there were three post-exposure measures of explicit word knowledge (form recognition, meaning recognition, meaning recall). New to this area of vocabulary research, outcome items were presented in randomized item modality (visual or auditory), to ensure congruence between treatment and test items and reducing modality-specific testing bias (Jelani & Boers, 2018). Group-level comparisons examined differences in (1) developmental trajectory of form familiarity and meaning integration for RO and RWL groups, (2) learning outcomes, and (3) effects of multi-componential L2 proficiency and phonological short-term memory (PSTM) skills on processing and learning outcomes. Within-RWL analyses operationalized a theoretical source of benefit (reading slightly ahead of the audio) and its impact on reading time and posttest learning gains.

Findings indicated differences between RO and RWL across three measures of eye movements: (1) gaze duration (GD), a measure of form familiarity with new words; (2) total reading time (TRT), a measure of meaning integration; and (3) visit count, or the total number of encounters looking at the words. The overall pattern for RWL indicated longer initial reading times for new words, fewer re-readings, and steadier decrease in GD and TRT across encounters.

Additionally, differences in learning outcomes were most clearly revealed through auditory test items, with RWL superior to RO across all three posttest outcome measures, and a group by item modality interaction. In other words, RWL indicated superior overall effects compared with RO across all items in form recognition and meaning recall, across all three posttests in auditory items, and better scores on visual than auditory items in RO (but equal across test item modality in RWL). Within-RWL analyses revealed that reading ahead of the audio was a positive predictor of TRT, as well as the most difficult of the three outcome measures (meaning recall). While PSTM predicted processing of new words, it did not predict outcomes for any of the three measures of vocabulary learning gains for advanced-level L2 readers. In sum, this study provides convergent evidence that process (form-form / form-meaning acquisition) and *product* (learning gains) are both positively impacted for new words under multimodal incidental conditions for advanced L2 learners, along with an initial indication that audiovisual asynchrony may play a role in RWL benefits in learning new words above and beyond L2 proficiency or memory skills.

TOWARD A THEORY-BASED ACCOUNT OF THE L2 VOCABULARY
PROCESSING AND LEARNING BENEFITS OF READING WHILE LISTENING

by

Jonathan Edward Malone

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:

Professor Kira Gor, Co-Chair

Assistant Professor Bronson Hui, Co-Chair

Dr. Nick Pandza

Dr. Stefanie Kuchinsky

Professor Donald Bolger, Dean's Representative

© Copyright by
Jonathan Edward Malone
2024

Dedication

To the rest of the Maryland Malones for all of our laughter, tears, joys, and sorrows on this adventure.

Acknowledgements

When one takes 10 years to complete a degree program, completion begets reflection. As I've been reflecting on the journey, there are many who need proper recognition, far beyond what I'm capable of attempting here. But here goes.

It was a great privilege to take my first course as an Advanced Special Student at UMD (still love the acronym) with Mike Long. He oozed love for SLA. Taking his course was like drinking from a fire hydrant of information, and I still think I convinced him to admit me into the program by introducing him to animations in a PowerPoint slide presentation. His reference letter for me was one of the funniest moments of my life through its unintentional comedy. He left an indelible mark on my work – I will always hear his voice when I'm tempted to split an infinitive, when I hear about CLIL, and when I think about unobtrusive input enhancement. His love for ISLA is what convinced me to do the Ph.D. here, and suffer through the training in experimental research. I'll never regret it. Thank you, Mike. I'll always remember your FC Barcelona flag draped over your desk, your two-finger typing style, your passion for methodological rigor in study design, your curiosity at every event I saw you at, and inviting me to share warmed-up Trader Joe's frozen quiche with you. Your warmth and hospitality were unprecedented for me to enjoy and now emulate in the academic world. I'll never forget how excited you were at my success. May your influence on the field remain strong in encouraging methodological rigor and a desire to know more about how L2 acquisition proceeds, and can be supported in better ways.

It was a gift to me to work with Kira Gor through the past seven years in the program. Kira helped me navigate the transition from my area of comfort in research (applied learning studies) and into the shark-infested theoretical waters of phonology and lexical representations,

which informed my second qualifying paper and this study. Kira was unfailingly patient and kind to me through years of developing materials, moving research remote during the pandemic, and then ceding some control over my studies when I wanted to do eye tracking and vocabulary learning for my dissertation study. Though I often knew how much I was adding to her burden as she helped lead the program through tumultuous years, it was never because she implied it. Kira, thank you for modeling consistency, mentorship, and resilience.

Navigating through the initial stages of the UMD program in experimental research required much patience on the part of the academic mentors I had, and I'm deeply grateful for the many hours spent unpacking and developing ideas in courses and for my QPs with Robert DeKeyser and Steve Ross. Robert's blend of intensity and genuine encouragement was motivating, and his help in assisting my understanding of individual differences was invaluable. The way he conducted his Ph.D. seminars was exemplary, and I structure many of the courses I teach in similar ways. And, of course, the Leonidas chocolate. Thank you, Robert. Steve was readily available when I had a question, and I'll always remember the odd DOS-era programs he would share with me to use for specific types of analyses. His patient explanations of confirmatory ANOVA, IRT, the centrality of reliability to measurement, and the role of variance in analyses helped me grasp the ideas. Trips to his mint green office, with its reclining chairs and mellow vibes, usually resulted in me having many more questions than answers when I left, but his work and words were always thought-provoking. Thank you, Steve. Your presence in the program is sorely missed, but your impact is still seen in your students.

It was a delight to me when the UMD program hired a fellow eye-tracking researcher in 2022, and even more so that it was a colleague. Bronson Hui had already impacted my work in considering issues related to methodology and measurement, but having him as a co-Chair on

this project has been a pleasure. Bronson's drive to collaborate imbues his research and mentorship with real and lasting relationships, and his growing reputation within the field as a stellar researcher is paralleled by a genuine interest in the development of his students. Thank you, Bronson, for modeling balance, good humor, and excellence. Your commitment to research and confidence in your work is matched by real humility, which draws others in and sets an example for your colleagues and students.

There are so many others to thank at UMD. Laura Stapleton taught me about histograms and distributions with a dash of dryly sardonic humor, while Kaiwen Man patiently unpacked inferential stats for me early in the program. Eric Pelzl, Payman Vafae, and Ilina Stojanovska were peer shepherds for me, helping me navigate the early days of impostor syndrome. Since then, it's been a parade of wonderful peer student researchers. Thank you Atsushi, Basak, Sunhee, Man, Beth, Catherine, Wei, Kichan, Matt, Ryo, Yingzhao, Fatima, and all of my other classmates. Thank you, Kyoko, for all of the lunches, Japanese gifts for our boys, and being willing to come and babysit when Caleb was born. Thank you, Ilaria, for working with me to understand how unobtrusive input enhancement can be meaningfully operationalized. Thank you to the current group, especially Meghan for fantastic voice work, Mireia for carrying the torch on incidental conditions, So-Hye for asking so many good questions, and the best possible eye-tracking data collection team I could have imagined. Tanya, Mona, Jill, Mireia, and Sanshiroh are the real MVPs of this study. Thank you for your consistency and good humor through the tedium of powering through 120 full participants along with me in eye-tracking data collection. What an effort. I hope none of you hate Tolstoy from now on.

To my MEI colleagues, thank you for everything. To Liz Driver, thank you for taking a chance on hiring me when I probably wasn't really ready, your encouragement to me through the

years, and your modeling of curiosity and readiness to learn. To Ray Smith, thank you for being generous with your time and energy, and being patient with my many administrative oversights and missteps. (The Danish butter cookies helped me get through this year). Your friendship and collegiality have been deeply meaningful to me. To Jennifer Moore, thank you for modeling hard work, determination, and consistency. Thank you to the many teachers we've had through the years, for showing me better what a strong skill-focused language teacher should (and sometimes shouldn't) be like in the classroom. The connections between my research and work have made this journey truly enjoyable, so thanks are also in order for the parade of students I've worked with through the years at MEI. I'm so much more impressed with you than me.

Gratitude also belongs with the grant support providers for this project, from the *Language Learning* journal community to Duolingo and the International Research Foundation for English Language Education. Your support made this project possible, and enabled me to train some excellent researchers in utilizing methods that were new to us, while also enabling me to compensate participants in a much more respectable way than is typical for a university-based research study. I don't take that for granted, and appreciate the support. I trust that those funds did not go to waste.

Thank you to Dad and Mom, who have always stoked my curiosity about the world, and have been dogged in trying to understand what I'm doing. Your love and generosity to me is an undeserved gift. Thank you Andy, Amanda, Sarah, Josh, Grant, and Mark. It's been a delightful ride to be siblings with you, and our closeness amidst diversity is a real joy. To our CHBC and CBC families, thank you for welcoming us to this area many years ago, making us at home far from home, and loving us with great joy as we travel together as sojourners in a strange land, seeking a new and better country. Thank you Abram, Caleb, Luke, and Seth. You've opened

vistas of joy, pride, and happiness in my life. Thank you for being patient with your absent-minded Dad, especially in the last couple of years. I could not be prouder or more grateful for each of you.

To My Dear Wife, I had no idea what this journey would entail when it started over a decade ago. Thank you for holding the Light when I've been a lost boy out in the trees. Thank you for being strong when I've been weak. Thank you for never pretending to be everything, and in so doing being everything I've needed for a spouse, lover, and COO (principal, instructor, choir director, piano instructor, and Mom) of our crew and our home. Thank you for laying down your life, time, and desires for the good of others, and for me. May we always find our lives individually and together by joyfully laying them down.

All remaining errors in this work are my own.

SDG

Table of Contents

Dedication	ii
Acknowledgements.....	iii
Table of Contents.....	viii
List of Tables	xi
List of Figures	xiii
List of Abbreviations	xiv
Chapter 1: Introduction.....	1
Chapter 2: Literature Review.....	5
2.1 Introduction.....	5
2.2 L2 Vocabulary Learning: Magnitude and Centrality.....	6
2.3 L2 Lexical Development: What is the Goal?.....	7
2.4 L2 Lexical Development in Sequence: Form and Meaning Mapping.....	12
2.5 Multimodality Facilitating Form-Form and Form-Meaning Connections.....	19
2.6 The Conditions for Learning Vocabulary	26
2.6.1 <i>Incidental vs. intentional conditions: apples and oranges</i>	27
2.6.2 <i>Lexical development under incidental conditions: exposure frequency</i>	30
2.6.3 <i>Lexical development under incidental conditions: modifying the input</i>	33
2.6.4 <i>Multimodality as input modification under incidental conditions</i>	35
2.7 The Process of Learning is Important, along with the Product.....	41
2.8 Measurement and Modality in Multimodal Designs	49
2.9 Individual Difference Factors in Learning Vocabulary under Incidental Conditions	50
2.10 Overall Summary of Literature Review.....	53
2.11 The Present Study	54
2.12 Research Questions.....	55
2.13 Initial Predictions	55
Chapter 3: Methods.....	62
3.1 Introduction to Methods.....	62
3.2 Overview of the Research Design.....	62
3.2.1 <i>Participants</i>	63
3.2.2 <i>Experimental task</i>	68
3.2.3 <i>Target item selection and pseudoword characteristics</i>	70
3.2.4 <i>Vocabulary posttests</i>	73
3.2.5 <i>Individual difference measures</i>	75
3.2.6 <i>Procedure</i>	78
3.2.7 <i>Apparatus</i>	81
3.2.8 <i>Auditory recordings and speed</i>	83
3.3 Data Analysis	84

3.3.1	<i>Eye-tracking data pre-processing and cleaning</i>	84
3.3.2	<i>Individual difference variables</i>	85
3.3.3	<i>Eye-tracking measures as outcomes for between-group analyses</i>	88
3.3.4	<i>Vocabulary learning outcomes for between-group analyses</i>	91
3.3.5	<i>Reading ahead of the audio as a predictor of TRT and VL in RWL</i>	92
3.3.6	<i>Qualitative survey data coding</i>	93
3.4	Summary of Methods.....	93
Chapter 4:	Results	100
4.1	Introduction to Results.....	100
4.2	Individual Difference Descriptives	100
4.3	Story Comprehension Scores	101
4.4	Reliability Estimates	102
4.5	Eye-Tracking Measures	103
4.5.1	<i>Graphs and descriptive statistics</i>	103
4.5.2	<i>Inferential statistics for between-groups eye-tracking data comparisons</i>	111
4.5.3	<i>Summary of results for between-groups eye-tracking comparisons</i>	117
4.6	Vocabulary Learning Outcomes	120
4.6.1	<i>Graphs and descriptive statistics</i>	120
4.6.2	<i>Inferential statistics for between-group VL outcome comparisons</i>	126
4.6.3	<i>Summary of results for between-group VL outcome comparisons</i>	132
4.7	Within-RWL Analyses for Reading Ahead of the Audio	135
4.7.1	<i>Reading ahead as a predictor of TRT</i>	135
4.7.2	<i>Reading ahead as a predictor of learning outcomes</i>	136
4.7.3	<i>Summary of results for within-RWL analyses</i>	139
4.8	Effects of PSTM on Learning Outcomes.....	140
4.9	Qualitative Debriefing Survey Analysis	141
4.10	Supplemental Analyses	145
4.10.1	<i>Eye-tracking measures and ID variable supplemental analyses</i>	145
4.10.2	<i>Learning outcomes and ID variable supplemental analyses</i>	148
4.10.3	<i>Reading ahead of the audio in RWL and ID variable supplemental analyses</i>	151
4.11	General Summary of Results	154
Chapter 5:	Discussion	155
5.1	Introduction to Discussion	155
5.2	Multimodal Reading and L2 Lexical Development.....	155
5.2.1	<i>The process of reading ± audio: stability, trajectory, and change over time</i>	155
5.2.2	<i>The products of reading ± audio</i>	167
5.2.3	<i>Reading ahead of the audio as a mechanism for RWL benefits</i>	172
5.3	Individual Differences in Lexical Processing and Vocabulary Learning	176

5.4 Descriptions of the RO and RWL Experience	178
5.5 Limitations	179
5.6 Implications of the Findings	184
5.6.1 Overall implications.....	184
5.6.2 Pedagogical implications.....	185
5.6.3 Research design implications	187
5.7 Future Directions	188
5.8 Conclusion	190
Appendices.....	192
Appendix A: Experimental Text for Treatment.....	192
Appendix B: Reading Comprehension Questions	202
Appendix C: Characteristics of Target Pseudowords	204
Appendix D: Form Recognition Test Items and Format	206
Appendix E: Meaning Recall Test Items and Format.....	207
Appendix F: Meaning Recognition Test Items and Format.....	209
Appendix G: Cloze Proficiency Measure	211
Appendix H: LexTALE Proficiency Measure	214
Appendix I: Reading Speed Proficiency Covariate	216
Appendix J: Language Background Questionnaire	217
Appendix K: Debriefing Survey Questions	218
Appendix L: Model Comparisons for Inferential Analyses.....	219
References.....	222

List of Tables

Table 1 <i>Designs and Findings for Parallel Studies of L2 Lexical Development under Incidental Conditions</i>	47
Table 2 <i>Initial RQs, Predictions, Tasks, and Measures in the Present Study</i>	61
Table 3 <i>Operationalization of All Variables in the Present Study</i>	62
Table 4 <i>Demographic Information of Participants</i>	67
Table 5 <i>Language Background of Participants</i>	68
Table 6 <i>Study Tasks and Procedure in Chronological Sequence</i>	80
Table 7 <i>Pearson Correlation Coefficients between Proficiency, Memory, and Processing Variables</i>	86
Table 8 <i>Statistical Models Utilized in the Study by RQ and Prediction</i>	95
Table 9 <i>Descriptive Statistics for Individual Difference Variables in the Study</i>	101
Table 10 <i>Reliability Estimates for Each Vocabulary Learning Outcome and Covariate Measure</i>	102
Table 11 <i>Means, Standard Deviations, and 95% Confidence Intervals for Gaze Duration by Group</i>	106
Table 12 <i>Means, Standard Deviations, and 95% Confidence Intervals for Visit Count by Group</i>	108
Table 13 <i>Means, Standard Deviations, and 95% Confidence Intervals for Total Reading Time by Group</i>	110
Table 14 <i>Correlation Coefficients for Relationships between Eye-Tracking Variables</i>	111
Table 15 <i>Best-fitting Model for Gaze Duration</i>	112
Table 16 <i>Best-fitting Model for Total Reading Time</i>	114
Table 17 <i>Best-fitting Negative Binomial Logistic Mixed-Effects Model for Visit Count</i>	116
Table 18 <i>Summary of Predictions and Results from Eye-tracking Measures</i>	118
Table 19 <i>Descriptive Statistics for Form Recognition Raw Learning Gains</i>	120
Table 20 <i>Descriptive Statistics for Meaning Recognition Raw Learning Gains</i>	122
Table 21 <i>Descriptive Statistics for Meaning Recall Raw Learning Gains</i>	124
Table 22 <i>Best-fitting Model for Form Recognition Outcome</i>	127
Table 23 <i>Best-fitting Model for Meaning Recognition Outcome</i>	129
Table 24 <i>Best-fitting Model for Meaning Recall Outcome</i>	131
Table 25 <i>Summary of Results from Vocabulary Learning Outcomes</i>	133
Table 26 <i>Best-fitting Model for RWL Analysis of Readahead Variable as Predictor of Total Reading Time</i>	136
Table 27 <i>Best-fitting Model for RWL Analysis of Readahead Variable on Form Recognition Outcome</i>	137

Table 28 <i>Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recognition Outcome</i>	138
Table 29 <i>Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recall Outcome</i>	139
Table 30 <i>Summary of Results from Analyses of Reading Ahead of the Audio</i>	140
Table 31 <i>Best-fitting Model for Gaze Duration (ID variables separated)</i>	146
Table 32 <i>Best-fitting Model for Total Reading Time (ID variables separated)</i>	146
Table 33 <i>Best-fitting Model for Visit Count (ID variables separated)</i>	147
Table 34 <i>Best-fitting Model for Form Recognition Outcome (ID variables separated)</i>	149
Table 35 <i>Best-fitting Model for Meaning Recognition Outcome (ID variables separated)</i>	149
Table 36 <i>Best-fitting Model for Meaning Recall Outcome (ID variables separated)</i>	150
Table 37 <i>Best-fitting Model for RWL Analysis of Readahead Variable as Predictor of Total Reading Time (ID variables separated)</i>	151
Table 38 <i>Best-fitting Model for RWL Analysis of Readahead Variable on Form Recognition Outcome (ID variables separated)</i>	152
Table 39 <i>Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recognition Outcome (ID variables separated)</i>	153
Table 40 <i>Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recall Outcome (ID variables separated)</i>	153
Table 41 <i>Research Questions, Predictions, and Reported Findings in the Present Study</i>	156
Table 42 <i>Lexical and Contextual Characteristics of Target Items in the Study</i>	204
Table 43 <i>LexTALE Proficiency Measure Items and Correct Answers</i>	214
Table 44 <i>Model Comparisons for Each Eye-tracking Measure for Group Comparisons</i>	219
Table 45 <i>Model Comparisons for Each Vocabulary Learning Outcome</i>	220
Table 46 <i>Model Comparisons for Readahead Variable Regressed on Outcomes within RWL Group</i>	221

List of Figures

<i>Figure 1.</i> Contextual information mean rating trendlines for each target word context by exposure.	73
<i>Figure 2.</i> Correlation coefficients and graphs for the three individual difference covariates.	88
<i>Figure 3.</i> Data visualization for gaze duration ET measure by group. (first plot is means; second plot is fitted loess lines with 95% confidence intervals.	105
<i>Figure 4.</i> Data visualization for visit count ET measure by group. (first plot is means; second plot is fitted loess lines with 95% confidence intervals).	107
<i>Figure 5.</i> Data visualization for total reading time ET measure by group. (first plot is means; second plot is fitted loess lines with 95% confidence intervals).	109
<i>Figure 6.</i> Best-fitting model-based estimates of gaze duration by group across instances.	113
<i>Figure 7.</i> Best-fitting model-based estimates of TRT across instances.	115
<i>Figure 8.</i> Best-fitting model-based estimates of visit count across instances.	117
<i>Figure 9.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on form recognition visual items.	121
<i>Figure 10.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on form recognition auditory items.	121
<i>Figure 11.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on all form recognition items.	122
<i>Figure 12.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recognition visual items.	123
<i>Figure 13.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recognition auditory items.	123
<i>Figure 14.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on all meaning recognition items.	124
<i>Figure 15.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recall visual items.	125
<i>Figure 16.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recall auditory items.	125
<i>Figure 17.</i> Bee swarm plot of mean score percentages and 95% confidence intervals by group on all meaning recall items.	126
<i>Figure 18.</i> Group by item modality interaction for the form recognition outcome. .	128
<i>Figure 19.</i> Group by item modality interaction for the meaning recognition outcome.	130
<i>Figure 20.</i> Group by item modality interaction for the meaning recall outcome.	132
<i>Figures 21-23.</i> Group and sample proportions of responses to dichotomous debriefing survey items.	142

List of Abbreviations

ET	Eye tracking
GD	Gaze duration eye tracking measure
ISLA	Instructed Second Language Acquisition
L1	A learner's first or native language
L2	A learner's second or subsequent language
PSTM	Phonological short-term memory
RO	Reading only
RWL	Reading while Listening
SLA	Second Language Acquisition
TRT	Total reading time eye tracking measure
VL	Vocabulary learning
WM	Working memory

Chapter 1: Introduction

Vocabulary knowledge plays a central role in both receptive and productive second language (L2) use (e.g., Zareva, 2005; Webb & Nation, 2017; Qian, 2002; Vafaei & Suzuki, 2020; Hui & Godfroid, 2021; Zhang & Zhang, 2022; Uchihara & Clenton, 2023, among others), and it is unsurprising that it remains a central aspect of second language acquisition (SLA) and instructed SLA (ISLA) research. A growing body of lab and classroom-based studies examining different learning treatments involving single-word lexical items has shed substantial light into how learning proceeds, and practical ways in which to support learners, both in decontextualized and intentional word learning tasks (see Webb et al., 2020 for a meta-analysis) and contextualized, meaning-focused tasks (see Uchihara et al., 2019 for one recent meta-analysis). Within contextualized learning designs, in particular, learning treatments under incidental conditions have typically reported modest learning gains in knowledge of new vocabulary, both on immediate and delayed posttests of form and meaning knowledge (see Tuzcu, 2023, for a recent discussion). These designs provide less-obtrusive ways in which learners gain access to information about the form and meaning of new words, while remaining focused on meaning in L2 tasks (see Doughty, 2008; Long, 2017).

Advances in technology have concurrently provided ways to diversify the way new L2 vocabulary are encountered under incidental conditions (both in context and modality), as well as more precise methods for measuring cognitive processing during and learning gains from L2 meaning-focused tasks. The rich linguistic environment of the digital language learning world, which now pervades even physical language classrooms, allows for an ideal location for testing predictions regarding multiple input modalities and their impact on learning. Learners encounter several input modalities immediately upon opening any language learning app and most

websites, with rich and diverse content available to instructors who may differ widely on teaching methodology, instructional design, and assessment. Simultaneously, the continued development of methods for assessing real-time language processing, especially related to reading and tracking eye movements as indicators of cognitive attention processes (Conklin et al., 2018; Godfroid, 2020a; 2020b), has provided new veins of research into how learners encounter new words during meaning-focused tasks, gain familiarity with their forms, and begin to integrate meanings in new lexical representations.

Substantial parallel research has focused on underlying psycholinguistic representations of L2 words, and relationships between types of linguistic information in lexical representations and their activation during language processing (e.g., Bordag et al., 2022). A growing body of evidence from L2 studies focused on lexical representations indicates dynamic interplay of form encoding of phonological, orthographic, and semantic information in a new L2 word's developing representation (e.g., Jiang, 2000; 2021; Gor et al., 2021; Bordag et al., 2022). Remarkably few empirical studies on new word learning, however, have tied input manipulations designed to maximize learning with underlying psycholinguistic theoretical motivations for interventions. Even fewer have made specific testable predictions regarding a mechanism for the benefits of these manipulations, and their implications for pedagogical practice and materials.

The present study was designed with both of these goals in mind. This dissertation study examines the relative benefits of one sub-area of multimodality, reading while listening (RWL), comparing it with processing and learning outcomes from reading only (RO) while predicting and testing a mechanism for previously-reported benefits of RWL on L2 vocabulary learning outcomes (e.g., Brown et al., 2008; Malone, 2018; Teng, 2018; Chen, 2021). To do so, I utilized eye-tracking (ET) methodology to examine real-time processing behavior, comparing groups in

three eye-tracking measures (gaze duration, total reading time, visit count), posttest vocabulary learning outcomes in three areas (form recognition, meaning recognition, meaning recall) and in multiple item modalities (visual, auditory), while accounting for multi-componential constructs of L2 proficiency, phonological short-term memory, and general processing ability.

This dissertation is organized into five chapters, including the current introduction as Chapter 1. In Chapter 2, a survey of relevant literature on theoretical matters related to the study is provided, centering on the task of vocabulary learning in SLA, how general SLA theories of attention and learning have impacted task and study design, challenges that prior studies have encountered and articulated (or not), the need for continued growth in examining real-time processing during learning, and the theoretical motivation for predicting and testing learning benefits from RWL. The present dissertation study is here situated within other recent work examining processing and learning outcomes for L2 vocabulary, with the specific research questions and predictions made at the point of proposing the study articulated.

In Chapter 3, the methods utilized in the present dissertation study are outlined, specifying how variables were operationalized, participant demographics and recruitment procedures, the experimental story task and its contextual manipulation, the item characteristics of target pseudowords, vocabulary posttests, individual difference measures, procedure employed, ET apparatus and procedures, information about audio recording, and an overview of data analysis procedures for all research questions and variables. In Chapter 4, results for each research question are provided and whether they confirmed predictions related to ET measures, learning outcomes, and individual differences, as well as a summary of qualitative data coding results regarding participant awareness during the learning task. In Chapter 5, the findings are summarized in light of my predictions at the time of the proposal, and discussed within broader

lexical development frameworks, as well as findings from other similar studies. Theoretical and pedagogical implications of the study are addressed, while acknowledging its limitations. I conclude with a summary of the findings, as well as directions for future research.

Chapter 2: Literature Review

2.1 Introduction

This chapter provides a narrative overview of the relevant research literature for the present dissertation study: L2 vocabulary development and learning under incidental conditions, L2 lexical processing, input manipulations, congruence between treatment and test items in assessing new word knowledge, and the impact of individual differences in proficiency and memory on both processing and learning vocabulary. First, I situate the study in light of the magnitude and centrality of the vocabulary learning task in an L2, and how previous research has not often connected well with broader theoretical discussions within SLA. Secondly, I discuss theoretical underpinnings and empirical evidence for lexical processing and representations in the L2 to proceed from form-form to form-meaning mappings, and how previous findings impact how contextualized word learning under incidental conditions may or may not proceed. This discussion will address broader theoretical and methodological issues regarding how the type of L2 knowledge can impact its deployment in L2 use, how theories of L1 and L2 reading impact how L2 vocabulary development may occur over time, and studies comparing different types of conditions and input modifications.

I then move to address empirical research on how multimodality in learning conditions may facilitate form-form and form-meaning connections, while acknowledging the challenges that multiple streams of input can provide to the learning task. I specifically situate the present study within designs that have included comparisons of RO and RWL groups. Problems with incongruence between treatment groups and test items during assessment across the literature are addressed, and how this discrepancy may explain some of the lack of reported benefits of multimodality on learning. Finally, the impact of individual differences in L2 proficiency and

phonological short-term memory (PSTM) on findings regarding vocabulary learning, both under intentional and incidental conditions, is discussed. At the end of the chapter, I present the research questions that motivated the study, along with initial predictions.

2.2 L2 Vocabulary Learning: Magnitude and Centrality

Vocabulary learning research has played a crucial role in SLA and Applied Linguistics fields over the past four to five decades (see Nation, 2006; Webb & Nation, 2017). A rich and diverse body of work has explored the effects of explicit teaching and learning of individual lexical items, stemming from long-held assumptions of the centrality of vocabulary to L2 development that extend further into the past than the relatively-recent field of SLA—philologists in the late 19th century recognized the centrality of lexical knowledge in L2 development, and reading ability specifically (e.g., Sweet, 1899). More recent arguments have further elucidated the magnitude of the vocabulary learning task, as learners encounter far more vocabulary items in the L2 environment than are possible to study explicitly inside a language classroom, yet are often highly successful in the task (Schmitt & Schmitt, 2012; Long, 2017). Conservative estimates have ranged from receptive knowledge of 11,000 words for Dutch university students (Hazenburg & Hulstijn, 1996), to 9,000+ word families (base form plus all related forms) in English as necessary for reading comprehension at an academic level (Nation, 2006; Nation & Chung, 2009).

While the task of L2 vocabulary learning is daunting for the learner in any language, its importance is relatively non-controversial. The relationship between vocabulary knowledge and components of L2 proficiency is strong, with connections reported between vocabulary knowledge and global L2 proficiency (e.g., Zareva, 2005), listening (Staehr, 2008; Milton et al., 2010; Vafae & Suzuki, 2020; Hui & Godfroid, 2021; Zhang & Zhang, 2022) and reading

comprehension (Qian, 2002; Nation, 2006; Stæhr, 2008; Zhang & Zhang, 2022). Much of this learning necessarily occurs outside the scope of explicit classroom focus, given time constraints in both second and foreign language learning contexts (Long, 2017; Webb & Nation, 2017). Therefore, there has been great interest into methods for helping learners study vocabulary independently (e.g., Webb & Nation, 2017). However, the sheer volume of words necessary for a high level of L2 proficiency can be intimidating, and reducing vocabulary learning to decontextualized activities (e.g., flash cards, dictionary definitions, or word lists) can risk oversimplification, especially when learning polysemous words (e.g., Webb & Nation, 2017). Decontextualized vocabulary learning, while useful in its own right, cannot account either for the size of the task or the way in which lexical development is nuanced by natural variation in semantic context and complexity (Hamrick & Pandža, 2020). Therefore, it is unsurprising that much parallel research in SLA has focused on manipulating contextualized, meaning-focused input to maximize lexical development even while the learner is focused on another task (see Long, 2017; Borro, 2021).

2.3 L2 Lexical Development: What is the Goal?

An important issue within applied and contextualized L2 vocabulary research to this point has been the so-called “black box” problem of L2 classroom research, and of ISLA in general, first outlined by Long (1980). In an attempt to preserve ecological validity, researchers have mostly set up large-scale input modifications of one type or another, set the class or experimental session in motion, and then measured “learning” through explicit, form-focused, written, and usually decontextualized posttests (for a recent example, see Barcroft, 2015). This input/output approach to ISLA research with vocabulary learning is understandable, given budgetary constraints and ecological challenges to classroom research, but is quite limited. It has

flattened the learning process and outcome into a pragmatic, simplistic, but easy to measure definition, often written L2-L1 translation, at the expense of methodology carefully grounded in psycholinguistic and SLA theory. Such studies cannot account well for the complexity and variety of the developing L2 lexicon (Jiang, 2000), lexical developmental patterns from form-to form to form-meaning connections (Chaudron, 1985; Webb, 2007; Bordag et al., 2022; Gor et al., 2021), or relationships between processing and outcomes (Godfroid, 2020a).

If the goal of lexical development is identifying a word's meaning in a multiple-choice test, or L1↔L2 translation from a list, research that only measures posttest learning outcomes from treatments is sufficient for making comparisons between instructional interventions. However, I agree with others (e.g., Godfroid, 2020a) who suggest that we need a more principled approach based on theoretically-driven predictions for lexical development, and evidenced by empirical findings regarding the development and representation of lexical items in the L2, nuanced by more complex methods for assessment of vocabulary knowledge. Additionally, literature into vocabulary learning has not often addressed broader discussions in SLA research of how acquisition and automatization proceed. Fluent, accurate, and automatic language processing and use have long been held as a central goal of SLA (e.g., DeKeyser, 2000; Segalowitz, 2008; Godfroid, 2020a), but have been conspicuously absent from measurements in L2 vocabulary learning studies. Indeed, as Schmitt (2010) noted, the lack of a comprehensive theory of vocabulary acquisition, and its relationship to L2 development, is deeply problematic for motivating a unified research agenda in vocabulary development.

The goal of L2 lexical development, as the field of SLA research defines it, should be both more nuanced and less exclusively dependent on explicit, form-focused knowledge that can be articulated on a paper test or a computer screen (whether receptively or productively). Rather,

L2 lexical development should be operationalized more broadly to include growth in rapid and automatic deployment of lexical knowledge in meaning-focused contexts, regardless of the type of skill in which it is deployed (see Webb & Nation, 2017), alongside traditional measures of learning. While more pragmatic in nature than certain broader theories of SLA might appreciate, this definition is larger in the sense that integration of new knowledge into the L2 lexicon can be contextualized, articulated, and assessed more systematically and comprehensively than simply through the traditional dichotomy of yes/no in word knowledge, operationalized through explicit tests. The more rudimentary tool of explicit posttests, while still an essential metric, cannot account for the dynamic process of knowledge development for a novel lexical item, especially in the establishment of its representation in short and long-term memory and the initial form-form and form-meaning relationships between L1/L2 words and concepts (e.g., Jiang, 2000; Bordag et al., 2021; Gor et al., 2021).

As Godfroid (2020a) noted, reliance on offline, explicit posttests severs the assessment of linguistic knowledge from real-life use, “often characterized by a certain amount of time pressure” (p. 434). Unsurprisingly, Godfroid (2020a; 2020b) has argued that the field of vocabulary research must account for online, procedural language processing (Ullman, 2004; 2006) as a complement and a predictor of offline outcomes. Given its increased efficiency and focus on meaning, Long (2017) similarly argued that a central tenet of the ISLA research agenda must focus on instructional interventions using meaning-focused real L2 language tasks, while triangulating measures of online processing and offline knowledge.

Including the measurement of online language processing during vocabulary learning from meaning-focused tasks has at least two benefits. First, it provides a window into the learning environment that more closely approximates naturalistic conditions. For proficient L1

readers, their encounters with new words almost exclusively occur within reading and/or listening and/or multimedia contexts, and are frequent. Therefore, there is an inherent ecological validity in L2 research with contextualized input through meaning-focused tasks, along with measurement of behavior in these types of conditions, in mirroring continued L1 lexical development. The goal of vocabulary acquisition and knowledge is rapid processing and efficient use in varied semantic contexts, whether in L1 or L2, and can better be delivered through contextualized input.

Secondly, measuring online processing of new vocabulary also allows for a closer examination of learner attention, and its distinct and contested relationship in the SLA literature with detection of the input, noticing, conscious awareness, and learner uptake (e.g., contrast Robinson, 2002, and Schmidt, 2012, with Tomlin & Villa, 1994). L2 lexical processing studies utilizing reaction time measures indicate for single word recognition that L2 learners access some semantic information very quickly (e.g., Li et al., 2017). However, as Godfroid (2020b) noted, reaction time task responses are particularly sensitive to reactivity effects, and variability within and between participants in reaction times. The addition of a behavioral task in the response impacts the quality and reliability of the measure as an indicator of the phenomenon, which limits the predictive validity of these types of studies and reported findings as to what lexical/semantic information learners are truly accessing.

It is highly likely that new vocabulary items, particularly salient forms that are fixated on for longer times such as meaning-bearing parts of speech (e.g., nouns, verbs, adjectives, adverbs) are noticed quickly in the input, and learners try to figure out their meanings from the context (see Godfroid, 2020a). This interpretation aligns with Ellis's (1994; 2008) account of attention being necessary for the learning of form, but both attention and awareness necessary for the

mapping of form and meaning. However, the distinctions between noticing / detection, implicit / explicit learning and knowledge, and conscious awareness are probably more fruitful at a theoretical level for less-salient forms, such as syntactic patterns (e.g., Leung & Williams, 2011; Rebuschat & Williams, 2012). Indeed, there seems to be growing evidence that statistical and implicit learning play an important role in linguistic development of less salient L2 features, such as multiword lexical patterns and grammatical features (e.g., Borro, 2021; Ren & Wang, 2023). These discussions are beyond the scope of the present study.

The examination of online behavior through monitoring eye movements has provided a new window into behavior during learning, and may provide important insights into the attention, detection, noticing, and awareness conversation. However, given the amount of time L2 readers process new words, and how quickly form familiarity develops across encounters with new words (e.g., Pellicer-Sánchez, 2016), it is both unlikely that processing for meaning remains implicit for long, and nearly impossible to measure (as Rebuschat & Williams, 2012, note). For the present study, it was important to include questions about awareness of the target items in the debriefing survey, in order to get a rough measure of the extent to which the new words were noticed, and what participants were aware of doing in order to learn word meanings. However, given the notorious difficulty both of operationalizing and measuring implicit learning and knowledge of vocabulary (e.g., Bordag et al., 2015), I do not make strong claims to be measuring implicit learning or knowledge of new word forms, much less meanings. In fact, debriefing questions indicated that participants were very aware of the new words, and consciously trying to integrate their meanings into their contexts during the reading task.

However the learning is defined, it is clear that the development of lexical knowledge can attain to rapid access and retrieval, especially during online processing in tasks such as reading

(Elgort & Warren, 2014), whether the outcome is situated within broader frameworks of implicit (Tomlin & Villa, 1994), procedural (Ullman, 2004; 2006), or automatized explicit knowledge (DeKeyser, 2008). Given the clear evidence of awareness of participants on the debriefing survey in the pilot for the present study, and the relative length of time and number of fixations that L2 learners exhibit with new words that other parallel studies have reported (e.g., Godfroid et al., 2013; Pellicer-Sánchez, 2016), I assume that much of the learning of meaning is explicit, especially during later encounters with new words, although it is certainly possible that the learning of form at early stages would not involve awareness. I am uncertain as to whether it will ever be possible to measure this lack of awareness.

The ultimate goal of ISLA research, language instructors, and learners is the same when considering learning under incidental conditions: facilitating fast, automatic, accurate, and fluent knowledge both of form and meaning. Developments in online monitoring of the learning process during meaning-focused activities through measures such as eye tracking (Conklin et al., 2018; Godfroid, 2020b) have provided new ways of opening the “black box” and examining its contents at a more granular level, which warranted the present study. Monitoring eye movements provide unprecedented access to real-time L2 processing and additional insights into cognitive mechanisms underlying development, especially as differences in conditions and/or behavior predict and exhibit superior learning outcomes.

2.4 L2 Lexical Development in Sequence: Form and Meaning Mapping

In many ways, ISLA vocabulary research has lacked clear ties to theoretical accounts of L2 lexical access and development related to psycholinguistic processes in which to motivate study design and pedagogical treatments. The most commonly referenced framework is from Nation (2001) and Nation and Meara (2013), who focused on sub-component parts and specific

subskills involved in learning new vocabulary. Their framework is useful for language learners and practitioners in considering teaching and learning, given their focus in these works on influencing vocabulary learning program-level decisions. However, these studies and others in the sub-area have not grounded vocabulary learning processes well alongside systematic psycholinguistic accounts for how L2 word learning proceeds, or how representation of new word knowledge is conceptualized, in motivating research study design. This is especially true given the differences between lexical networks emerging in the L2 from the L1, with the two almost certainly overlapping to some degree during language development (Jiang, 2000). Models of L2 lexical access and development (e.g., Jiang, 2000; Gor et al., 2021; Bordag et al., 2022) have included the core tenet that while L1 lexical representations are robust, complex, and accessed quickly and easily (Hulstijn, 2001), the quality of the development of form and meaning encoding (i.e., phonology and/or orthography and/or semantics) of L2 lexical representations proceeds on a continuum. This encoding is influenced by L1 factors such as cognate status or L2-specific phonological and orthographic information, and how L2 information relates to semantic information within specific L2 contexts (see Darcy, 2022).

Given perceptual and articulatory constraints on late L2 learners (Long, 1990; Granena & Long, 2013; Flege & Bohn, 2021), difficulties in encoding L2 phonological information in particular can facilitate problematic or “fuzzy” L2 phonolexical representations (Gor et al., 2021), resulting in frequent semantic confusions both in auditory processing (Carney, 2021) and in visual word recognition (Ota et al., 2009), even for *familiar* L2 words. Bordag et al.’s (2022) Ontogenesis Model (OM) provides a road map for lexical development at the item level, with a particular representation moving along a continuum of knowledge at separable but related phonological, orthographic, and semantic levels. Even in the lively debate among the 17

accompanying articles to the OM's publication which were published in the *Bilingualism: Language and Cognition* journal alongside it, none questioned the centrality of these three sources of information in the developing L2 lexical representation. Consequently, study designs including orthographic and/or phonological information in a semantic context are urgently needed, including treatments that facilitate the encoding of form and meaning in the L2 lexicon within naturalistic learning tasks. However, much of this work with semantic confusions has remained in auditory processing to this point (Ota et al., 2009; 2010 are exceptions). As discussed below, multimodal conditions provide an optimal context for examining development that includes simultaneous phonological, orthographic, and semantic information.

Theoretical models of reading, both in L1 and L2, have reflected similar assumptions about how both word decoding and semantic integration proceed by utilizing phonological information. Given the strength of a child's phonemic representations in the L1, in general, it is assumed that the graphemes and phonemes are mapped well during typical L1 reading development in alphabetic languages, especially in L1s marked by orthographic transparency such as Arabic or Spanish (see Frost et al., 1987). However, learners from logographic or syllabic languages have also been found to access phonological information while reading (e.g., Wang et al., 2003). Additionally, when L2 readers from transparent L1s begin learning to read in L2s with more complex relationships between phonology and orthography (such as English – see Harm & Seidenberg, 2004; Coltheart, 2005), problems generally ensue. L2 readers, especially those with weaker or developing L2 phonemic awareness, have greater difficulty with such mappings, especially when there is substantial difference in L2 orthographic information compared with L1 (e.g., script differences, level of transparency, script type) (see Wang et al., 2003; Ziegler & Goswami, 2006; Ziegler et al., 2010; Grainger & Ziegler, 2011; Brysbaert,

2022). When there is an overreliance on “bottom-up” processing, in which orthographic processing is mediated by phonological processing, inaccurate representations of L2 sounds and symbols can compound, leading to slower and more inefficient reading (see Hui, 2024, for an overview).

In L2s such as English, with relatively deep orthographic structure, visual word decoding has been hypothesized to proceed in one of two ways, depending on the nature of the lexical item and its representation in episodic memory. Word identification proceeds either in a direct print-to-speech, or lexical route, at higher orthographic grain sizes and in a top-down fashion (see Coltheart, 2005; Grainger & Ziegler, 2011), or through the mediation of sublexical phonology, through a bottom-up decoding process (Brysbaert, 2022). It is generally thought that in reading orthographically opaque languages such as English, the ideal state for most words is to be read at the whole-word lexical level, or at least in larger chunks than phonemes. Regardless of whether positions on the development of word recognition in English come from serial (e.g., Coltheart, 2005) or connectionist (e.g., Harm & Seidenberg, 2004) frameworks of visual word recognition, both agree that efficient decoding and processing proceed from strong encoding of orthographic, phonological, and semantic information. Early and weaker L2 readers are often more dependent on a sublexical or phonological processing route in reading in English (see Brysbaert, 2022), which can be more taxing and inefficient when L2 readers cannot reliably depend on L2 phonological representations during reading, especially when encountering new words. As such, it is crucial to find ways to facilitate form encoding by strengthening ties between orthographic, phonological, and semantic information during initial stages of vocabulary learning.

According to Perfetti’s (2007) *lexical quality hypothesis* for L1 English reading, higher resolution of form-form connections allows for faster and more accurate lexical identification

and retrieval (see also Elgort et al., 2018 in an L2 context). Thus, the development of lexical representations in English during reading is both simple and complex. Simple, in that it involves the triangulation of three types of information about a word – its sound (phonological), symbol (orthographic), and meaning (semantic). Complex, in that L2 lexical development does not proceed in a linear fashion, with information better encoded for a given representation at a given time than for another, potentially even related word. Additionally, phonological and/or orthographic encoding may be more or less robust at the level of the word itself (Gor et al., 2021; Bordag et al., 2022). A number of recent studies have chronicled the effects of fuzzy L2 lexical representations (Cook et al., 2016; Darcy & Thomas, 2019; Llompert & Reinisch, 2019; Gor & Cook, 2020; Gor et al., 2021) on fast and accurate auditory L2 processing, and robust effects of the mutual interdependence between these three types of lexical encoding during L2 lexical development. These findings suggest that triangulating the type of encoding during instructional interventions regarding lexical development could prove useful in strengthening bidirectional form-form and form-meaning links during learning, in auditory and visual word decoding and semantic association (Ota et al., 2009; 2010).

Given the earlier discussion of the centrality of form acquisition at early stages of L2 development in general (Chaudron, 1985), and in form-form relationships throughout L2 lexical development (e.g., Bordag et al., 2022), it is clear that form acquisition plays an essential role both in establishing a novel word in the L2 lexicon, and also facilitating faster decoding during reading, allowing for deeper and better processing of the surrounding context and a wider variety of contexts in which the word occurs. Widely-accepted models of both L1 and L2 reading make a theoretical connection between eye movements, cognitive control, and the mind (e.g., EZ-Reader from Pollatsek et al., 2006; Cop et al., 2015). The EZ-Reader model assumes, based on

empirical findings, that reading in both L1 and L2 proceeds in a serial fashion, involving both orthographic and phonological processing. According to Pollatsek et al. (2006), “in the task of reading, one attempts to decode print in order to produce a representation of a spoken utterance that is logically sequential in nature” (p. 9). In other words, phonological and orthographic information are concurrently active during decoding and processing individual words, and also more broadly in higher-order semantic integration processes within a sentence to construct a meaningful whole. Again, the connections between phonological, orthographic, and semantic information during linguistic tasks are revealed.

The decoding process works remarkably quickly, both in the L1 and L2, but is much slower overall in the L2, even in reading familiar words (e.g., Conklin et al., 2020). The second and related assumption of EZ-Reader is that initial processing time in looking at words during fluent reading involves what is called the *familiarity check*, wherein gaze is fixated on a new word briefly, followed by a mental trigger that moves the eyes forward to the next word. This process is considered both minimal and efficient; Rayner et al. (2012) state, “we think it makes some sense for the system to do ‘good enough processing’ to direct eye movements forward and rely on other clean-up operations when something goes wrong” (p. 163). When the initial familiarity check is interrupted, as in the case of an unfamiliar word or irregular form that is detected, the eyes are directed to remain looking at the information, and sometimes backtracking within the context, to construct the logical explanation for the interruption. Under this simplified explanation for how eye movements and mental processing proceed, a variety of semantic contexts when encountering information about a new word would be extremely useful to higher-level semantic integration of novel word information (see Hamrick & Pandža, 2020).

With the prominence both of L2 processing and cognitive control models of reading such as EZ-Reader, it is unsurprising that arguments surrounding L2 vocabulary acquisition and learning from a theoretical perspective have assumed the centrality of initial form-form, as well as form-meaning connections (Wesche & Paribakht, 1996; Webb, 2007; Barcroft, 2009). The establishment of strong connections between phonological, orthographic, and semantic information would be paramount in such a framework. If the initial step of familiarity check is essential to fluent word processing, as well as higher-level semantic integration, it should be of central interest to research. Until recently, however, studies have been unable to examine the developmental trajectory of the familiarity check (i.e., the initial pass through an interest area including a word target), as few in L2 vocabulary research have monitored eye movements during reading. However, it is now possible to do so, as several recent studies have done (e.g., Godfroid et al., 2013; 2018; Pellicer-Sanchez, 2016; Elgort & Warren, 2014; Elgort et al., 2018; Tuzcu, 2023).

These studies have utilized *gaze duration* (GD) as a proxy for the familiarity check, or the strength of form-form connections, across a time course of multiple instances of exposure to new words in context. GD is the total duration of all eye fixations in a target area during the first pass through the interest area, usually defined in reading research around a single target word of interest. This aligns with the argument from the EZ-Reader model that the familiarity check is an efficient and simple method of triggering eye movements through familiarity with form (e.g., Cop et al., 2015). If eye movements are interrupted, and gaze duration is lengthened during initial encounters with a new word, this indicates that the reader has encountered an interruption in the flow of form recognition. Subsequent encounters with the new words should indicate decreasing GD times, as familiarity with form and speed of recognition increase. Re-visiting the

interest area including the new word is then thought to indicate broader challenges with initial stages of meaning integration within the context (see Godfroid, 2020b).

2.5 Multimodality Facilitating Form-Form and Form-Meaning Connections

The building argument from a theoretical perspective is that there is emerging evidence that L2 lexical development proceeds from initial form-form connections, namely a new word's phonology (sound) and its orthography (symbol(s)). Mapping the two appears to be a key aspect of learning to recognize individual words, no matter the theoretical perspective on their relative contributions (e.g., Harm & Seidenberg, 2004; Coltheart, 2005), as well as being central to continuous word processing during reading (Rayner et al., 2012). As such, recent innovations in online processing measurement have made it possible to test theoretical accounts regarding the real-time behavior of L2 learners as form-form and form-meaning links are created and strengthened in real time (see Godfroid, 2020a and Hui, 2021 for overviews of measurement). However, relatively few have examined online processing during presentation of simultaneous input modalities during linguistic tasks, and subsequent effects on specific learning outcomes. This is surprising, given the growing attention given to multimedia in research, and the proliferation of language learning applications that utilize multiple input modalities.

L2 studies that have included online measurement of eye movements while utilizing multimodal conditions have been few in number, and have included comparisons between L1 and L2 reading behavior for familiar words during reading while listening tasks (e.g., Conklin et al., 2020), how much learners look at subtitles when watching videos (e.g., Peters & Webb, 2018), and gaze patterns when simultaneous audio, text, and images are presented (e.g., Serrano & Pellicer-Sanchez, 2019). To my knowledge, only one study (Tuzcu, 2023) directly compared differences in reading patterns for unfamiliar words among L2 learners in RO and RWL groups

through tracking eye movement patterns, with the purpose of comparing the process of how form-form links are established under incidental conditions. Chen (2021) utilized an online semantic priming task to explore deeper knowledge of form and semantic encoding in learning new rare English words from RO and RWL groups, but did not find group effects for priming of new words. However, the brevity of the learning task in that study (a maximum of four exposures to target words) and its posttest nature likely made measuring early traces of such learning very difficult. Developments in more sensitive and non-reactive online assessment such as eye tracking have made such comparisons more feasible, both in early and late measures of eye tracking behavior (Conklin et al., 2018; Godfroid, 2020). We can now observe behavior under incidental conditions that isolates effects of multimodal conditions on learning processes and outcomes.

The psycholinguistic and methodological benefits of RWL are potentially twofold. First, reading while listening provides an unobtrusive pathway to helping L2 readers establish form-form connections, by providing a direct and simultaneous comparison between visual text (orthography) and auditory information (phonology) during a continuous and fluent stream of linguistic information. This is particularly helpful for L2 learners from L1s that are orthographically transparent in learning words that are phonologically opaque (e.g., *through* or *city*), but would facilitate connections between orthography and phonology for learners from any L1 background by providing accurate L2 phonological exemplars in the input (see Uchihara et al., 2022). In both cases, this would facilitate processing at larger orthographic grain sizes across exposures to words through accelerating form-form mapping in a bidirectional way, which is essential in efficient word recognition in languages such as English (see Grainger & Ziegler,

2011), while facilitating phonological and orthographic “bootstrapping” (Gor et al., 2021) in visual word identification during early repeated encounters with new words.

Secondly, RWL could provide statistical learning opportunities and repetition effects by duplicating and diversifying form exposure, allowing for additive effects of exposure frequency beyond visual occurrences in a text. There is a growing consensus regarding the robustness of contextual exposure frequency effects relating to vocabulary learning outcomes, over and above corpus-based word frequency reported in multiple recent studies of lexical access (e.g., Adelman et al., 2006 for L1; Chen et al., 2018 & Hamrick & Pandža, 2020 for L2), as well as statistical meta-analyses of L2 vocabulary learning (Uchihara et al., 2019; Yanagisawa & Webb, 2021). Duplication of exposure in varied contexts through multiple verbal channels (visual/auditory) is therefore predicted to facilitate form-form and form-meaning connections. The natural ties between orthographic and phonological information that simultaneous presentation provide, coupled with a rich and diverse semantic environment, should facilitate automatization of form knowledge and subsequently deeper form-meaning connections.

This position is not without tension. Bisson and colleagues (2013; 2014; 2015) have argued elegantly for lexical development to benefit from simultaneous modalities by situating their studies within Paivio’s Dual Coding Theory (Paivio & Csapo, 1973), wherein both verbal and nonverbal information are encoded simultaneously when multiple modalities (e.g., verbal linguistic information and pictorial information) are presented. Kinchla’s (1974) Redundant Signals Effect, in which processing of novel nonverbal information such as lights/tones or letters/sonar targets was found to be better when multiple types of information were presented simultaneously, was expanded into a linguistic context by Lewandowski and Kobus (1993). They found that word recall of new words among L1 speakers was significantly better when presented

through an audiovisual bimodal channel than when presented in either mode alone. Similarly, Montali and Lewandowski (1996) extended these findings to reading comprehension with less-skilled L1 readers.

However, as Mayer and colleagues (2001) and others have argued, presenting multiple input modalities can be taxing on human cognition. Moreno and Mayer (2002) replicated Lewandowski and Kobus (1993) with L1 speakers in comparing reading while listening comprehension outcomes with reading only, but when additional visual material was added (e.g., visual animation), these effects disappeared. They interpreted these findings to indicate that at a certain level, the cognitive load induced by multiple input sources reached its maximum threshold, resulting in decreased cognitive benefits (Sweller, 1988; 2011; Kalyuga et al., 1998). Additions to Cognitive Load Theory (Mayer et al., 2001; Mayer & Johnson, 2008) have extrapolated that these ideas extend to RWL, not just input sources with non-verbal information (as Mayer and colleagues initially argued), with studies such as Diao and Sweller (2007) indicating that text comprehension among L2 learners was found to be helped less by RWL compared with RO. However, this study included very short passages, and was read at an astonishingly slow pace (76 words per minute). It was conducted with first-year English majors in China, who had substantial experience learning English. Even conservative estimates assume that fluent L2 reading typically proceeds at a much faster pace than 76 words per minute (Suk, 2017), so it is likely that the L2 readers in Diao and Sweller (2007) read ahead much faster than the audio, nullifying or mitigating potential benefits that would come from synchrony or near-synchrony of audio with the reading. This possibility was not considered in the Diao and Sweller (2007) study, which did not examine where participants were looking during the RWL task.

A meta-analysis by Adesope and Nesbit (2012) summarized and reported effects from 20 studies comparing RWL comprehension outcomes with RO or listening only, with no images or animation. Of these 20 studies, only two reported negative effects, and one of the two was Diao and Sweller (2007) mentioned above for its methodological problems. Most studies in Adesope and Nesbit's (2012) meta-analysis reported robust and positive effects of RWL on comprehension outcomes, although that comparison was not often with RO groups. While other recent studies (e.g., Hui, 2024) have indicated that comprehension may not be significantly facilitated by RO, none of these studies have shown superior effects of RO over RWL in comprehension outcomes. At the very least, it appears that RWL does not detract from comprehension, while potentially providing benefits in other ways.

Since studies into cognitive load have not included online processing measures, as L2 studies reporting benefits of RWL have utilized (e.g., Serrano and Pellicer-Sánchez, 2019), reports of deleterious effects of multiple input sources may be a side effect of a lack of relative synchrony of reading with audio, rather than overtaxing mental capacity. In other words, participants may be either reading or listening, while ignoring the other source of input. An important insight into real-time L1 and L2 reading behavior was reported in Conklin et al. (2020), who compared L1 and L2 eye movement patterns of reading behavior in RO and RWL groups to measure differences among university-level readers. The text consisted of two 1,500-word experimental stories, and the audio was recorded at a rate of roughly 210 words per minute – slightly slower than Goldman-Eisler (1961) reported for typical native speaker speech rate in English, but nearly three times faster than the audio in Diao and Sweller (2007).

Interestingly, Conklin et al. (2020) found that reading was infrequently synchronized with the audio, albeit much more often for the L2 group (17% for L1; 33% for L2), and that for

the asynchronous samples, a very high proportion (nearly 90% for the L1 group, and nearly 80% for the L2 group) reflected reading slightly ahead of the audio. In the case of lexical processing, if learners are able to read slightly ahead of the audio, while keeping it in immediate proximity, the phonological decoding process and establishment of phonological-orthographic form-form connections would benefit substantially from immediate auditory feedback. In other words, providing the audio could facilitate an immediate hypothesis test about the pronunciation of a newly-encountered visual word, allowing for duplicative effects and better encoding of it into memory, and strengthening the fledgling initial memory trace even when the primary focus of the activity is on broader comprehension.

At this point, I should briefly engage the work of what has become the companion piece to this dissertation, Tuzcu (2023), who compared online reading behavior in RO and RWL for 63 advanced-proficiency English learners reading a 9,700-word selection of a novel under incidental conditions. The author reported differences in reading pattern behavior between RO and RWL groups in GD (considered an index of the familiarity check), as participants in the RWL group consistently read novel pseudowords longer during initial encounters. However, there were no significant differences in TRT, usually associated with higher-level semantic integration, at the group level. Tuzcu (2023) also reported that participants in both groups exhibited incremental learning gains in target pseudoword items, with significant group effects (RWL > RO) both on immediate and delayed form recognition posttests, and higher mean scores but no significant effects found for meaning recognition or meaning recall posttests. Additionally, Tuzcu (2023) found robust lexical learning effects through reading times in a sentence-reading posttest (RO > RWL in reading times), indicating a positive impact of RWL on multiple facets of lexical development.

Tuzcu's (2023) findings that GD times were longer in RWL were interpreted as evidence of RWL drawing learner attention to the targets for longer during initial encounters, increasing their salience (as Long, 2017 suggested). The sharper decrease in GD in RWL in that study across exposures indicated form familiarity developing more quickly in RWL than RO, facilitating faster processing during later exposures to target items in RWL than RO. However, there were two methodological choices in Tuzcu (2023) that differed from the present study. First, participants in the RO group were given as much time as they wanted to read each trial, and were told to push a button when they were ready to proceed to the next trial. Conversely, the participants in the RWL group read and listened to each trial, but they automatically proceeded to the next trial following the audio recording completing the trial. The author reported that participants in the RO group spent, on average, nearly a full second longer on each trial than participants in the RWL group (Mean RWL = 30,286 ms, $SD = 0$; Mean RO = 31,138ms, $SD = 10,769$ ms). In other words, the average summed reading time for all words (including the context surrounding the pseudowords) across trials was slightly over a full second longer in RO than RWL in this study. One possible implication of this methodological decision is at the level of semantic integration. The potential benefit in RO of having additional time on each trial may have influenced the meaning-based outcomes (no difference between RO and RWL in meaning recognition and recall), as participants in RO were able to spend more overall time integrating the text and the new word meanings into their unique semantic contexts. This additional time on task may have obscured effects of group (RO vs. RWL) on explicit outcomes if participants spent that time reading the text surrounding the targets, and could not be accounted for in statistical models.

Secondly, Tuzcu (2023) also made the deliberate decision not to give explicit instructions to participants in RWL to read along with the audio. While this is an understandable choice, and was done so to preserve additional ecological validity in the study, the author did not measure the alignment of reading patterns with the audio (see Conklin et al., 2020). The fact that participants may not have been well aligned with the audio, coupled with the additional time afforded to the RO group, may explain some of the lack of effects in that study regarding the pattern of reading novel pseudowords (which was largely identical in both groups) and the lack of effects on meaning-based outcomes, either in recognition or recall. These design-level decisions can have an impact on outcomes, especially since Tuzcu (2023) was the first study to my knowledge that actually compared real-time reading patterns in RO and RWL, and predicted learning outcomes from reading times. These types of differences in methodological choices are an important aspect of study design for vocabulary learning under incidental conditions, which makes it an imperative for study designs to specify these choices. It is worth spending some time at this point examining the general historical lack of consensus regarding such differences in operationalizing learning conditions in relation to other methodological choices, which can impact findings.

2.6 The *Conditions* for Learning Vocabulary

The magnitude of the task and the limited time learners have to study decontextualized word lists in developing lexical knowledge broadly implies that contextualized word learning is occurring, whether in the L1 or L2 (see Webb, 2017). While this process has been criticized as both slow and incremental (e.g., Nagy et al., 1985; Laufer, 2003), it logically proceeds from recognizing the absence of sufficient time for explicit, form-focused activities and exercises to account both for the breadth and depth of lexical knowledge necessary for fluent receptive and

productive word knowledge. Clearly, learning in naturalistic contexts is occurring, and often very successfully. In the following section, three primary areas of research into vocabulary learning under incidental conditions will be reviewed briefly, along with their direct implications for the design of the present dissertation study: (1) studies comparing incidental and intentional conditions; (2) frequency of exposure; and (3) manipulation of the input. Methodological choices in each of these areas must be accounted for if meaningful comparisons between findings are to be obtained.

2.6.1 Incidental vs. intentional conditions: apples and oranges

A wide range of studies in L2 vocabulary learning have pitted learning gains from intentional learning conditions (whether with or without context) directly against learning outcomes from conditions claimed to be incidental in nature (e.g., Hulstijn et al., 1996; Laufer, 2001; Barcroft, 2015). Such comparisons, while understandable from a pedagogical perspective, proceed from three substantial, interconnected, and problematic assumptions that have impacted methodological choices. First, these comparisons commit a logical either-or fallacy, in which the goal of the comparison is to find which is superior to the other in some way, without a critical examination of whether the comparison itself is warranted. There are vast cognitive, linguistic, and pedagogical differences between tasks that involve presenting words in lists to memorize, compared with presenting learners integrated text with instructions to focus on meaning. As such, vocabulary learning studies would be better employed in utilizing tasks and measuring learning either from intentional or incidental conditions, but not both, and without comparisons. Secondly, operational definitions of incidental conditions vary greatly, with some (e.g., Laufer, 2001) claiming that as long as learners were not originally intending to commit the vocabulary to memory, the learning itself is considered incidental, even if participants looked up definitions in

a dictionary during the linguistic task. Third, these types of studies often assume that the intentional and incidental learning conditions result in the same level of learning outcome, and measure accordingly. It is then unsurprising to see that task designs focused on explicit, form(s)-focused learning produce better explicit, form(s)-focused outcome scores, but this inherent bias is rarely acknowledged.

These three assumptions, largely springing from simplistic definitions of vocabulary learning and knowledge, likely resulted in spurious comparisons, with some declaring that explicit focus on lexical form and meaning is the “best” way to teach and learn new words (e.g., Laufer, 2003). Others have argued for the opposite position, that optimal L2 vocabulary development can and should proceed from context alone (e.g., McQuillan & Krashen, 2008; McQuillan, 2019). This disagreement over presuppositions, a sort of “chicken and egg” conversation regarding optimal L2 lexical development, has driven a wedge into the research into instructional design for supporting vocabulary acquisition and learning.

At this point, it seems clear that explicit study of vocabulary, whether studying word lists, flash cards, or stopping a reading or listening activity to look a word up from a dictionary, produces some benefit for adult L2 learners (e.g., Webb & Nation, 2017), who are largely more dependent on conscious and effortful L2 study (Bley-Vroman, 1988; DeKeyser, 2008). Effects on those types of outcomes are robust and clear, at least for the sorts of untimed and explicit tests used to measure resulting knowledge from them. However, if (as mentioned earlier) the goal is fast, fluent, and accurate processing of lexical information and rapid memory retrieval during contextualized semantic processing in real-world linguistic tasks, comparisons between outcomes from differing manipulations of incidental conditions are essential to examine how meaning-based input could be modified by instructors and curriculum designers to maximize

processing and learning efficiency among L2 learners, while also facilitating faster word identification and meaning retrieval in tasks that involve a focus on meaning (e.g., Long, 2017). Consequently, the present study is not designed to make claims regarding whether one should teach or learn vocabulary only under intentional and/or incidental conditions. Those comparisons are less useful until the picture of what actually occurs in the online processing arena during encounters with new words in meaning-focused tasks is better clarified.

As it relates to methodological decisions I made for the present study, this study was designed to examine the learning process for L2 learners when they are reading, and the way process can impact more traditional learning outcomes in multimodal learning situations. Explicit learning in decontextualized activities requires certain mechanisms for learning, and produces measurable knowledge. Contextualized learning conditions including a focus on meaning have typically resulted in gains in the same type of vocabulary knowledge, although reported gains are usually smaller (but perhaps better integrated through contextual semantic variety). What is crucial to the present study's design is how I focused on contextual learning under time pressure, where everything regarding the new word must be inferred quickly from the context. As I and others have argued elsewhere (Bruton et al., 2011; Malone, 2018), the term incidental needs careful definition, both in how it is conceptualized and operationalized.

Even with more advanced tools of examining online processing through monitoring eye movements, we are still making inferences about processes in the mind that are not well understood. As such, researchers are able to manipulate the conditions for learning, but probably not the learning itself. In other words, the term *incidental learning* may largely be a misnomer when discussing vocabulary development, particularly given the salience of new lexical items within a text. Models of both L1 and L2 reading typically assume that the reader is driven toward

comprehending the meaning of a text, and that learners move their eyes quickly to process and decode words for the sake of lexical access and integration. Given the speed of this process, especially among high-proficiency readers, research designs would be served better by avoiding the term incidental learning altogether, and instead clarify that designs are established to examine learning that occurs under incidental conditions. I have sought to do so in the present dissertation study.

2.6.2 Lexical development under incidental conditions: exposure frequency

A second primary vein of both classroom and experimental research into vocabulary learning under incidental conditions has focused on the relationship between exposure frequency and offline, explicit learning outcomes. Uchihara et al. (2019) conducted a meta-analysis of 26 studies, and 45 effect sizes, to isolate and summarize effects of frequency of exposure in the input on explicit vocabulary learning (VL) outcomes. They found that learner (individual differences in age and vocabulary knowledge), item (type of enhancement, engagement, variability in input), and methodological variables all contributed to variability in the size of such repetition effects, but that a medium ($r=.34$) effect size was found for repetition across studies. This finding of reported effects from studies under incidental conditions aligns with recent findings in psycholinguistic-based studies both in L1 and L2 that semantic and contextual diversity of experiences with lexical items contributed at least as much to speed of lexical access as corpus-based frequency, the traditional gold standard for lexical access (e.g., Hamrick & Pandža, 2020).

As I argued in Malone (2018), methodological differences are common across studies examining the effects of exposure frequency and utilizing incidental conditions, from differences in text type (experimental sentences, experimentally-manipulated passages, modified graded

readers, unmodified graded readers, chapters in a novel) to the use of words or pseudoword targets. Shorter texts flooded with instances of real-word targets typically resulted in greater gains across different types of word knowledge (e.g., Rott, 1999; Webb, 2007; Malone, 2018; Chen, 2021; Serrano, 2023), whereas designs utilizing longer texts and/or the use of pseudoword targets often report more modest learning gains (e.g., Chen & Truscott, 2010; Elgort & Warren, 2014; Godfroid et al., 2018).

As the definition of “learning” has not often been clearly specified in studies examining effects of exposure frequency, it has often been operationalized as the establishment of form-meaning links. However, even given methodological variation in these studies, results have indicated measurable learning of new words even when exposures are few (5-10), with the caveat that roughly 95% of the other words in the passage should be familiar for fluent reading (Waring & Takaki, 2003). Coupled with online processing findings that Pellicer-Sánchez (2016) and Godfroid et al. (2018) reported regarding more fine-grained measurement, a robust picture has emerged that initial stages of form-form and form-meaning links are acquired when there are 8-10 exposures with a new word in a text or a series of texts, even from conservative estimates (e.g., Rott, 1999; 2007; Brown et al., 2008). As such, results from studies of frequency of exposure generally narrow the scope of focus on the first 8-10 instances, as well, especially in cross-sectional and lab-based research studies (see Pellicer-Sánchez, 2016; Elgort et al., 2018; Godfroid et al., 2018; Tuzcu, 2023).

An additional complicating factor in study methodology regarding frequency of exposure effects in L2 vocabulary learning is that these studies often do not meaningfully control for the number of total encounters with a new word. Given individual differences in proficiency, vocabulary knowledge, and speed/fluency, a reader may return to an earlier word many times

if/when the new word is noticed, reading it over and over again intentionally within its context to derive its form and meaning. This is particularly true when no control is exerted over time on task, which has predominated in classroom research designs in particular. Unfortunately, many of the studies reporting learning outcomes under incidental conditions make confident assertions about a specific number of encounters with novel words without accounting for multiple encounters with a word within a single context during the meaning-focused reading task. This could result in inflated outcome scores for learning from additional unmeasured repetitions.

If researchers are to make claims regarding exposure frequency and vocabulary learning outcomes, the element of time pressure can help direct attention more to meaning in the task, allowing for clearer comparisons between declarative and procedural knowledge under more stringent incidental conditions (see the discussion on timed and untimed measurement in Suzuki, 2017). Additionally, more targeted assessment of total encounters with a word can provide a clearer picture on frequency of exposure effects. Time pressure in the current study was provided through automatic progression through the screens where the story was presented, with no possibility of returning to a previous screen, in both RO and RWL. While this type of reading activity is less ecologically valid, it allows for more specific claims about online processing including the time course variable of instance within the text than would be possible if there were no time pressure applied during the reading task. Additionally, frequency of exposure was measured in the current study through the visit count eye-tracking measure, or the total number of times a reader's eyes entered the interest area on the trial that included the target item. This metric sheds additional light onto how much participants are actually encountering the new words.

2.6.3 Lexical development under incidental conditions: modifying the input

The manipulation of linguistic input in order to enhance or optimize vocabulary learning outcomes through particular instructional interventions has been of foundational interest to the ISLA research enterprise (Doughty, 2008; Loewen, 2014; Long, 2017). Doughty and Williams (1998) produced a taxonomy of the degree to which input manipulation is more or less obtrusive to a meaning-focused task by drawing attention to form (see table on p. 258). While input flood (increasing/decreasing the degree and repetition of linguistic input) and *unobtrusive* input enhancement do not direct attention to form, activities such as dictogloss, dictionary definitions of novel words, and other consciousness-raising tasks were defined as *obtrusive* in nature, by redirecting attention of the learner away from meaning and onto linguistic form during an instructional or interventional activity. Doughty (2008) and many others agree that ISLA with adult L2 learners should include a mix of both, but the emphasis on more obtrusive versions of activities including attention to forms has predominated in ISLA research (see Long, 2017).

This distinction has manifested in studies of lexical development, even in work claiming to facilitate learning under incidental conditions, often involving attention-grabbing designs, and frequently based on Laufer and Hulstijn's (2001) Involvement Load Hypothesis. They argued that greater vocabulary learning (again defined simplistically as immediate and explicit uptake) proceeds from intentional and effortful tasks focused on learning new vocabulary. Studies in this vein involving obtrusive input enhancement during meaning-focused tasks involve conditions that are more neatly defined as *semi-incidental* in nature (Pellicer-Sánchez & Boers, 2018). Examples abound, on a spectrum of salience from typographic enhancement (Kim, 2006; see Lee & Huang, 2008, for an overview), marginal glosses (Hulstijn et al., 1996; Hulstijn & Laufer,

2001; Laufer, 2003; Kim, 2011), explicit word-writing tasks (Elgort et al., 2018), and even the use of dictionaries or provided definitions embedded within a text (Laufer, 2001; Akbulut, 2007).

If, however, the ultimate goal of instructional intervention in ISLA is maintaining a focus on meaning, with only the occasional direct and explicit intervention into the learning task with selective focus (as Long, 2017 suggests), designs and curriculum on the *unobtrusive* side of the input enhancement spectrum (Doughty, 2008) are optimal. Given the extent of the task, and the still-open question as to the relative involvement of slow and effortful processes compared with statistical sensitivity to input, adjustments to input without such dramatic disruption of focus on meaning have produced important findings. In vocabulary learning studies, this type of input modification has typically involved manipulating the distribution of novel words in meaning-based contexts (spaced vs. massed), lexical or textual elaboration, and multiple simultaneous input modalities (reading and/or listening and/or viewing).

Studies examining the distribution of novel words in a learning task, whether close together (massed) or spread apart (spaced), have provided some evidence that larger spacing in multiple semantic contexts produces better and more nuanced learning outcomes (Cepeda et al., 2008; Webb & Chang, 2014; see Brown, 2021 for a counterexample). In general, the argument has proceeded that a variety of contexts for encountering new words strengthens form-meaning links, especially for polysemous words with nuanced meanings (Webb, 2017). This makes sense, given the wide variety of contexts and differences of meaning for a word such as “bank” in English, whether *the earth near a river* or *a place in which to keep money*. Additional variability in context would add semantic texture and depth to a developing lexical representation in a more natural way than a static dictionary definition could. This argument aligns well with the psycholinguistic finding that lexicality of word knowledge is equally sensitive to repetitions in a

range of meaning-focused contexts as it is to corpus-based frequency metrics (Adelman et al., 2006 for L1; Chen et al., 2018, and Hamrick & Pandža, 2020 for L2). Developing knowledge of word meaning across contexts appears to facilitate deeper knowledge and faster deployment of that word knowledge.

Related to repetition within a variety of semantic contexts, unobtrusive lexical or textual elaboration has been a growing area of ISLA research on vocabulary outcomes (see Kobayashi Hillman, 2021, for an overview). These interventions have included elaborating the input at the lexical level (Chung, 1995) by providing synonyms for target words (e.g., Vidal, 2011; Nguyen & Boers, 2019) or appositive grammatical cues (Godfroid et al., 2013). Conversely, structural elaboration has largely involved repetition of the input (i.e., manipulating frequency of exposure), paraphrase, and logical connections to improve semantic flow of the text and integration of novel items into a text (Kobayashi Hillman, 2021). These types of input manipulation have consistently found modest gains in using minimally modified naturalistic texts, indicating that learners are sensitive to semantic context when they encounter new words, providing additional evidence that semantic context is an important factor in learning new words under incidental conditions.

2.6.4 Multimodality as input modification under incidental conditions

The final sub-area of this discussion surrounding input modification has come through the unobtrusive input enhancement provided by multiple input modalities. One result of the information revolution of the 1980s and 1990s on the SLA field has been interest in the role of learning from differing types of linguistic input, and through multimedia particularly (Mayer et al., 2001; Mayer & Fiorella, 2014); in fact, an entire recent issue of the *Studies in Second Language Acquisition* journal was devoted to the topic (see Montero Perez, 2020), and the

involvement of technology both in the SLA research endeavor and in instructional design in ISLA research was recently reviewed by Chun (2016). Accordingly, vocabulary researchers have viewed multimedia as a fruitful avenue of unobtrusive input enhancement, whether in comparing learning outcomes from input provided in single modalities with each other (reading vs. listening), single modalities compared with multiple (listening compared to RO compared to RWL), or in combining text with audio and/or images and/or video. Easy and inexpensive access to such input, coupled with its potential for benefit, makes it ideal for both experimental and classroom L2 comparisons. Given current theoretical perspectives on the importance of multiple sources of information during language processing and learning in general, and the development of lexical representations specifically, it is unsurprising that ISLA researchers have sought to examine effects of differing presentation modalities in instructional treatments, and their effects on contextualized vocabulary learning in particular.

Carbo (1978) argued that “talking books” were extremely useful for weaker L1 readers in developing literacy skills, from comprehension to fluency and phonological familiarity for new words. Blum et al. (1995) extended the argument to L2 learners, reporting that children with audio support in a repeated reading program were more fluent in reading than children who read the text without simultaneous audio. Similarly, Taguchi et al. (2016) argued that reading while listening facilitated L2 reading fluency by parsing continuous speech, providing perceptual “dividing points”, and allowing for immediate and automatic comparison between phonological information presented in the audio and the automatic phonological subvocalization involved in silent reading (see Rayner et al., 2012). In L2 vocabulary research, Bürki (2010) reported better learning gains during intentional vocabulary learning activities when audio was provided. 88 Korean EFL university students studied 62 novel vocabulary words through paired associates

activities, and the author found consistent and durable superior performance on learning outcomes when participants were provided both auditory and visual information about the new words. Brown et al. (2008) found highest explicit form-meaning vocabulary learning gains from reading under incidental conditions for RWL, compared with RO or listening alone (although the comparison did not reach statistical significance), in a multimodal learning study of small L2 intact classroom groups. In general, robust effects have held across learners at varying proficiency levels (e.g., Webb & Chang, 2012, for beginning L2 learners; Malone, 2018, for intermediate; Chen, 2021, for high intermediate; Tuzcu, 2023 in form recognition for advanced), and have extended to multiword units as well (e.g., Webb & Chang, 2020; Borro, 2021; see a recent summary in Uchihara et al., 2022).

Webb and Chang (2012) examined lexical development among 82 beginning-level L2 English learners in Taiwan through a longitudinal repeated-reading program. They found greater gains in the RWL over RO, moderated by reading skill and vocabulary size among participants, and argued that providing auditory support when reading can facilitate L2 lexical development during early stages of L2 proficiency among adult learners. Teng (2018) examined VL outcomes from RO and RWL, with 60 intermediate EFL learners in Taiwan reading a graded reader story of approximately 6,000 words. Posttest results indicated better learning of both form and form-meaning connections, as well as collocational knowledge, from RWL. Malone (2018) compared two groups of 40 intermediate L2 English learners, examining explicit VL outcomes from RO or RWL to four short stories, and found significant group effects on both form and meaning recognition outcomes after two and four exposures to rare English words. These outcomes were moderated by both proficiency and phonological short-term memory (PSTM) for form recognition, but were robust to proficiency and memory effects. Chen (2021) replicated

Malone's (2018) findings for explicit learning outcomes, but did not find evidence for lexicalization of the novel words through a semantic priming lexical decision posttest.

Two additional studies (Brown et al., 2008; Serrano, 2023) utilized similar designs to Webb and Chang (2012), operationalizing incidental conditions through extensive reading programs with college students in Japan and 10-11 year olds in Spain, respectively. Each examined intact classes to compare learning outcomes from RO and RWL across months of L2 development, and found noticeably higher gains from RWL than RO (as Webb & Chang, 2012, did). However, neither found the differences to be significant, which limited their claims of longitudinal benefits for RWL. However, both of these studies included very small sample sizes for each group, compared with Webb & Chang's (2012) sample, so emerging differences may have been difficult to detect. Additionally, as I will discuss below, learning gains based on phonological mapping were not assessed through auditory test items in any of these studies, so the pattern of superior learning from RWL may be much greater in reality than previous study designs could detect. In the parallel cross-sectional study to the present dissertation, Tuzcu (2023) found a nearly-identical pattern of superior scores from RWL across three learning outcomes (form recognition, meaning recognition, meaning recall) to Brown et al's (2008) and Serrano's (2023) patterns, with participants in the RWL group scoring noticeably higher on outcomes, but these differences only reaching statistical significance for the form recognition outcome. While Tuzcu's (2023) sample size (29 in RWL; 30 in RO) was closer to Webb & Chang's (2012) 41 participants per group with high school English learners in Taiwan, Tuzcu (2023) included only visual items on the learning outcomes. Again, this limitation could obscure learning benefits, especially in measuring developing phonological knowledge.

Uchihara et al. (2022) extended the pattern of findings regarding superior effects of RWL to the learning of pronunciation in an intentional vocabulary learning design, comparing outcomes from RO, RWL, and listening only for 75 Japanese learners of L2 English during a word-image association task. They found that spoken word form learning was stronger in RWL than listening only, and that participants were significantly better and had more accurate pronunciation of the new words in a picture-naming posttest in both modes that included auditory information (listening only and RWL) than the RO group. The authors interpreted these findings to indicate that providing simultaneous text and auditory support provided reinforcement of fledgling and developing L2 lexical representations at both orthographic and phonological levels. Importantly, audio-supported conditions were especially sensitive to sound-spelling consistency of the novel words, with audio support helpful in participants mapping sound and symbol bidirectionally for more phonologically-opaque forms. Crucially, in Uchihara et al. (2022), RWL provided the orthographic support that listening-only lacked, while giving the phonological support that RO lacked.

Other work has also utilized multiple input modalities that involve viewing videos and/or textual information through subtitles. Syodorenko (2010) examined written and aural vocabulary learning outcomes among 26 high beginner learners of Russian from three experimental groups (video + audio + captions, video + audio, video + captions). The groups with captions scored better on written recognition of novel word forms, and the group including audio, video, and captions (both visual and auditory modalities) was superior to either video + captions or video + audio on the form-meaning outcome. Syodorenko (2010) interpreted these results as running counter to Robinson (2003), who argued that multiple input sources could overtax L2 learners, and Sweller's (2011) cognitive load theory, which posited that presenting simultaneous

information in a redundant way negatively impacts information processing in the L1 (see earlier discussion on cognitive load). Rodgers and Webb (2020) reported similar results to Syodorenko (2010) in their larger-scale longitudinal study. In Rodgers and Webb (2020), participants in the experimental group watched ten episodes of an English-language TV series in addition to their EFL university coursework, while participants in the control group were simply given a textbook-based EFL course. Participants who watched the TV series learned significantly more words, measured in two tests of form-meaning connections, than learners in the control group, and these results held in a delayed posttest.

Additional recent findings in viewing videos with subtitles appear to support Syodorenko (2010) in examining lexical gains from multiple input sources. Peters and Webb (2018) examined VL under incidental conditions, by presenting 63 Flemish-L1 learners of L2 English with a full-length Business English TV program. Utilizing a pretest-posttest-delayed posttest design, Peters and Webb (2018) found that participants in the experimental group who watched the course performed significantly better on a form-meaning recall posttest than the control group. Additionally, they found that prior vocabulary knowledge, exposure frequency / repetition, and cognate status of the target word with an L1 word were all significant predictors of learning, with large effects reported. Muñoz et al. (2021) examined vocabulary and grammar learning among 39 Catalan-Spanish bilingual students learning L2 English through watching videos with either L2 captions or L1 subtitles. Results indicated no difference in VL outcomes between groups that received captions or subtitles, indicating that L2 subtitles provided similar support to L1 captions in establishing L2 form-meaning links.

Ultimately, while many of these studies were skillfully designed, well implemented, and produced interesting and meaningful results, relatively few have provided a theory-based account

for why simultaneous input modalities can and should benefit learners, in what way(s), and even fewer have made testable predictions for potential mechanisms driving these benefits. Many have suggested possible sources for these benefits. Some have argued that multiple modalities assists the mapping of graphemes to phonemes (Goswami & Bryant, 1990), L2 text and phrasal segmentation (Webb & Chang, 2015), and affective benefits that could impact motivation (Tragant & Vallbona, 2018). However, the empirical case for psycholinguistic benefits has been sparse, and even contested at times, with very few testable claims beyond the possible sources of benefit briefly mentioned in discussion sections of journal articles. Unfortunately, most of these suggestions have gone completely untested. In order to justify a research agenda focused on providing multiple input modalities as a form of unobtrusive input enhancement for supporting lexical development, the theoretical basis underlying assertions of the benefits of multimodality on processing must be matched by methodological rigor and consistency in examining both processing and learning. Additionally, accounts for why multimodal conditions result in better outcomes should be accompanied by empirical predictions regarding the source of these benefits, along with methods designed to test them.

2.7 The *Process* of Learning is Important, along with the *Product*

Recent SLA vocabulary research, armed with tools previously unavailable in examining real-time processing phenomena, has provided a promising pathway and new insights into the relationship between real-time contextualized encounters with new vocabulary, along with connections between online processing and offline learning outcomes (e.g., Godfroid et al., 2013; Pellicer-Sánchez, 2016; Godfroid et al., 2018; Elgort et al., 2018; Tuzcu, 2023). In general, these studies have examined the relationship between measured attention to new word forms and explicit word learning outcomes within a natural context such as reading, and used eye-tracking

measurement to map the development of form familiarity and predict offline learning gains. As discussed earlier, utilizing eye movement data as a proxy for attention during reading processes aligns with cognitive control models of attention during reading (e.g., Rayner, 1998; Reichle et al., 2003; 2013; Rayner et al., 2012). An important assumption of these models is a tight connection between cognitive mechanisms in the reader and attention, defined and measured by tracking both early and late eye movements (Godfroid, 2020b). In other words, the so-called eye-mind link (e.g., Rayner et al., 2012) is hypothesized to be very tight during sequential tasks such as contextualized reading, with the “spotlight” of eye fixations within foveal vision (<1 degree of visual angle from the eye fixation) primarily involved in lexical processing (e.g., Reichle et al., 2009). Computational models such as E-Z Reader (Pollatsek et al., 2006) have successfully modeled how eye movements indicate the process of lower-level conversion or decoding of visual information for reading in English into phonological and orthographic information, as well as semantic codes, to understand a word’s meaning in its given context. This remarkably complex process occurs very quickly during fluent reading, and involves additional layers such as oculomotor control and larger semantic integration processes (see Godfroid, 2020b for an overview). In general, behavioral patterns in L1 reading and L1-L2 differences are assumed to be involved in L2 reading, but the same assumption of a strong relationship between attention and processing during reading remains intact (see Conklin et al., 2018; Godfroid, 2020b).

Given the centrality of attention in SLA research, the relationship between online processing and offline outcomes is a natural area of research focus. In general, and unsurprisingly, recent work within vocabulary acquisition and learning regarding time spent attending to novel word forms has reported a relationship between late eye-tracking measures (primarily total reading time) and explicit vocabulary outcomes, especially in learning new word

meaning (Godfroid et al., 2013; Pellicer-Sánchez, 2016; Godfroid et al., 2018). As such, as Paribakht and Wesche (1999) argued, the form-meaning connection necessary to encode semantic information in memory may indeed require explicit focus and noticing, rendering so-called “incidental vocabulary learning” an irrelevant or insignificant term. In other words, all vocabulary learning, whether contextual or decontextualized, may be in large part intentional and deliberate. If vocabulary learning is narrowly defined as explicit form-meaning connection, measured by explicit form-meaning posttests, that would certainly be the case. However, some studies have also included measures of lexicalization through priming tasks (e.g., Elgort, 2011; Elgort & Warren, 2014), self-paced reading (Bordag et al., 2015), and even the online trajectory of eye-tracking measurement across exposures (Pellicer-Sánchez, 2016; Godfroid et al., 2018), to explore how quickly form-form links can be created with novel words and integrated into the mental lexicon, even when initial memory traces cannot yet be articulated explicitly. These insights can shed light on how fast, automatic processes can be developed in L2 lexical processing, whether that process begins to a greater or lesser degree of explicitness and/or awareness.

Table 1 summarizes four recent studies that have utilized eye tracking to measure the development of form familiarity for new words for L2 readers under incidental conditions, with a focus on text comprehension, and including both eye-tracking measures and offline learning tests. Each of these studies helped to shape the predictions of the present study. Pellicer-Sánchez (2016) was the first to map online L2 vocabulary development and form familiarity at the level of individual exposure across encounters with new words, with 25 L1 and 23 L2 participants reading a 2,300-word short story with six target nonwords embedded eight times each. Explicit outcomes followed the predicted pattern of form knowledge tests producing stronger effects than

meaning knowledge tests, but the innovative element of this study was its time-course mapping of individual GD (a measure of early lexical processing) and TRT (considered a measure of lexical integration into its semantic contexts) across eight encounters with novel pseudowords.

For both GD and TRT, Pellicer-Sánchez (2016) reported a decrease in time during the first four-five exposures to novel words, followed by a leveling out, with participants reading targets similarly to familiar control items by the seventh or eighth exposure. The author argued that participants were becoming familiar with novel word form around encounter four, and had integrated form-form connections in order to process contextual information for meaning by encounter eight. However, the TRT in this study could have been jeopardized by a lack of control over duration of exposure for each screen. The participants could not return to a previous screen, but they were able to spend as much time as they wanted on a single screen before choosing to proceed.

In a similar vein of work, Godfroid et al. (2018) compared 19 L1 and 35 L2 English speakers' processing times of novel Dari words from reading five chapters (roughly 9,000 words) in an authentic English novel. They found a nearly identical S-shaped processing curve to the results reported in Pellicer-Sánchez (2016), as results of growth curve modeling reflected a non-linear pattern of decrease in TRT across initial encounters until the four-five exposure range, followed by leveling off and more gradual decrease at later exposures. While variation between the three target words of interest existed, the pattern of results held. Godfroid et al. (2018) also found relationships between reading time and explicit vocabulary posttests, which the authors interpreted as possible evidence of implicit, or at least increasingly automatic, knowledge of form:

[T]he present findings suggest that changes in eye movements over time can reveal word learning processes that are not detected by traditional measures of explicit vocabulary knowledge...[s]pecifically, the speed-up over time could reflect implicit learning processes or the gradual build-up and specification of a new word representation that can be accessed increasingly fluently during reading. (p. 574).

Elgort et al. (2018) utilized a very similar design, but with an expository text and rare English words, examining eye movements from various early (first fixation duration; GD) and late (go-past time; TRT; fixation count; regression count) eye tracking measures, as well as learning gains in sentence reading and meaning generation posttests. They reported a significant decrease in the difference between the measures for new words and familiar word controls by the eighth exposure in the text, which they interpreted from the early eye tracking measures to indicate that the familiarity check for new words was well-established by that point. However, orthographic familiarity and meaning integration for the unfamiliar words still took longer to occur for new words than familiar controls even after dozens of encounters in the text, which the authors took to mean that orthographic and semantic integration is a process that takes much more time than studies examining new traces of new word knowledge can fully map.

Tuzcu (2023), summarized earlier, extended these findings in comparing RWL and RO groups to examine relative differences both in processing trajectory for early (GD; regression path duration) and late (rereading duration; TRT) eye-tracking measures, as well as learning outcomes. The study examined processing and outcomes for 63 advanced English learners from a variety of L1s, who read a ~9,700-word selection of a graded reader version of H.G. Wells' *The Time Machine* embedded with 1-16 instances of novel pseudoword replacements for nouns, verbs, and adjectives. The author reported a significantly faster decrease in GD and regression

path duration across instances in the text for the RWL group than the RO group, whereas there was an equal and significant decrease in rereading times and TRT for both groups across instances in the treatment task.

These four studies suggest that the development of familiarity with new words across encounters appears to reflect automatization occurring in real time. Online processing measurement has provided new insights into how learners encounter new words in context during a meaning-focused task at each individual encounter, when learners may detect information about a word without consciously attending to its meaning (e.g., Tomlin & Villa, 1994). Meaning would then be constructed in-context with each encounter, moving from conscious and effortful inference to fast, fluent access (e.g., DeKeyser, 2008). Word form familiarity appears to develop very quickly and with less difficulty during reading for meaning, albeit differently than familiar words even after multiple exposures. This may result in making new words easy to recognize, but delaying automatic lexical access for word meanings until additional semantic processing occurs (e.g., Chaudron, 1985; Elgort et al., 2018; Chen, 2021). Since the field is just beginning to explore online lexical development in real time, these questions can be extended. Broadening the definition of learning to include form-form (phonological-orthographic) connections and form familiarity especially, as well as semantic encoding, is now a viable option for vocabulary research designs. However, it is highly important within these studies to control for prior knowledge of words, lexical characteristics, and item-level factors such as contextual information, or part of speech (e.g., Birch & Rayner, 2010).

Table 1*Designs and Findings for Parallel Studies of L2 Lexical Development under Incidental**Conditions*

Study	L2 participants (<i>n</i>) / L1s / L2 proficiency	Learning task	Eye tracking measures	Reported findings
Pellicer-Sánchez (2016)	25 L2 English learners / 11 different L1s / advanced proficiency	2,300-word short story embedded with six nonwords eight times	First fixation duration; gaze duration; total fixations; TRT	Gaze duration and TRT decreased significantly across the first 5-6 instances
Elgort et al. (2018)	34 Dutch-L1 English learners / high- intermediate to advanced L2 proficiency	Introduction and chapter (~12,000 words) of an expository text (<i>Freakonomics</i> by Levitt & Dubner, 2006) embedded with 14 rare words (ET analysis of the first eight exposures)	First fixation duration; gaze duration; go- past time; TRT; fixation counts; regression count	Novel pseudowords were read similarly to controls by the eighth exposure in the text; evidence of optimization of familiarity check by exposure 8; meaning integration (through late measures) remained different from novel pseudowords and controls throughout the text
Godfroid et al. (2018)	34 L2 English learners / a variety of L1 backgrounds / advanced L2 proficiency	Five chapters (~9,000 words) of an authentic novel in English	TRT	Significant decrease in TRT across exposures 1-10 for all three targets; TRT predicts form recognition learning outcome, but not meaning recognition or recall

Tuzcu (2023)	63 L2 English learners / a variety of L1 backgrounds / advanced L2 proficiency	Seven chapters (~9,700 words) of a graded reader version of a novel (<i>The Time Machine</i> by H.G. Wells)	Gaze duration; regression path duration; rereading duration; TRT	Significant decrease in gaze duration, regression path duration for RWL, but not RO group; significant decrease for rereading times and TRT that was equal in both groups
--------------	--	--	--	---

At this point, since it is now possible to track the development of form familiarity (through GD) and meaning integration processes (through TRT and regression or visit count data) in much more fine-grained ways in monitoring eye movements, a more nuanced and multi-dimensional definition of word learning is emerging. This definition includes the creation and strengthening of form-meaning links, but also the integration of psycholinguistic theory and language processing research alongside word learning and input manipulations. Unfortunately, sensitive measures of the learning of form familiarity and form-form (phonological-orthographic) links are not often utilized in ISLA vocabulary learning research, whether due to difficulty in assessment or because brief early memory traces of new word learning have been difficult to measure until recently (see Godfroid, 2020a). Given models of the dynamic nature of L2 lexical representation and development (Gor et al., 2021; Bordag et al., 2022), it is time for vocabulary researchers to account for these insights in ISLA research studies on the initial stages of vocabulary learning, as these four studies have done. One cannot relate a form to a meaning for a novel word if one cannot recognize or properly decode the form itself, whether at the level of phonological or orthographic information.

This is particularly relevant in an orthographically-complex L2 such as English, in which lower-level phonological decoding is widely thought to be involved in the initial processing of

unknown words (e.g., Coltheart, 2005). Given the necessity of larger-grained chunking and lexical encoding at larger grain sizes in English (e.g., Grainger & Ziegler, 2011), lower-level decoding soon gives way to easier access to orthographically opaque words through a direct lexical route as reading skill develops (Coltheart, 2005). Therefore, inefficiency in form-form mapping can adversely impact both the accuracy and rate at which form and meaning are linked, and through which whole-word lexical processing develops. As such, measures of both processing and learning of form are included in the present study as crucial indicators of lexical development.

2.8 Measurement and Modality in Multimodal Designs

An important point has been raised in a number of recent studies regarding L2 lexical development (Montero Perez et al., 2015; Hatami, 2017; Jelani & Boers, 2018; Uchihara et al., 2022) regarding comparative findings between groups utilizing multiple input modalities compared to a single modality, and how most previous work lacked treatment-test item congruence. These authors argued that previous studies nearly always utilized posttest items in the visual modality only as learning outcomes, which biased results toward conditions that include visual information, at the expense of phonological gains from learning conditions that included audio. As a result, they argued that future studies should be designed with such congruence in mind, and could reveal true differences in outcomes. The companion study to the present dissertation (Tuzcu, 2023) only included outcome items for form and meaning in the visual modality, so the present study extends those findings to include a multidimensional approach to assessment of pseudoword knowledge. As I have here argued for a more fully-orbed view of lexical access, assuming the mapping of phonological forms to be a crucial stage of

learning, it was essential that assessment of learning in the present study included phonology, and was operationalized through auditory test items included on the three VL outcomes.

2.9 Individual Difference Factors in Learning Vocabulary under Incidental Conditions

An often-underrepresented element of the existing body of research into lexical development in experimental contexts has been accounting for the influence of cognitive individual differences, or utilizing them as predictors of processing and learning outcomes. Despite extensive evidence even reported as early as Stanovich (1986) of bidirectional influences of cognitive abilities and higher-order processing during reading, most SLA research into vocabulary learning from context has given little more than cursory attention to measuring individual differences in proficiency, and accounting for them in models of processing and learning. While this is understandable, given time and budgetary constraints for cross-sectional designs, more comprehensive measurement of L2 proficiency is needed. It was particularly important in the present study to account for multidimensional aspects of proficiency, including both visual and auditory skills, given the focus on multimodal conditions and the need for construct validation within a given study. Otherwise, these individual differences could increase either Type I or Type II error of detecting group differences based on the input manipulation, depending on variability between and within participant groups. Since parallel studies in SLA include data from relatively few participants (typically 50 or fewer *total* in eye-tracking studies), at least some divergent findings regarding reported effects could be based on limited sampling and even more limited measurement of L2 proficiency, memory, and processing ability.

Along the same lines, there have been multiple recent findings regarding the connections between individual differences in language aptitude, and specifically working memory (e.g., Linck et al., 2013; Bonilla et al., 2020; Doughty & Mackey, 2021) and L2 learning ability in

general. There is some evidence that differences in PSTM, as defined by Baddeley (1998) involving both processing and storage components, impacts L2 lexical development. Kaushanskaya (2012) conducted two experiments, each with 18 bilingual English/Spanish speakers and 36 monolingual English speakers, and trained participants in phonologically-acceptable auditory nonwords paired with their English “translations.” The author reported that PSTM capacity was a significant predictor of learning outcomes for both bilingual and monolingual participants when learning pseudoword meanings.

Martin and Ellis (2012) trained 50 English-L1 participants in an artificial language learning study in an intentional learning design, and found strong correlations between vocabulary comprehension and production outcomes with PSTM scores on two nonword PSTM measures (nonword repetition and recognition). Elgort et al. (2018b) utilized a contextualized intentional learning design with novel vocabulary to examine effects of lexical elaboration, and included a measure of working memory (WM) (operation span, or Ospan) as a covariate. They found that complex WM significantly correlated with intentional VL outcomes, corroborating Martin and Ellis’s (2012) findings.

In the context of vocabulary learning under incidental conditions, Malone (2018) found that individual differences in composite working memory (WM) (Ospan and PSTM) were a significant predictor of outcomes on form recognition of rare L2 words for 80 intermediate-level L2 English learners, reporting an aptitude-by-treatment interaction (see Roehr, 2012) by input modality. Participants in the RWL group tapped into WM resources to a greater degree in form recognition than participants in the RO group during the first two to four exposures to new words. In Montero Perez (2020), 63 Dutch-L1 learners of L2 French watched a documentary in French, in which 15 French words were substituted with pseudowords, and were given form and

meaning recognition posttests of learning. Generalized estimating equations analyses indicated that complex WM tasks (operation span and backward digit span) were significant predictors of both form and meaning recognition outcomes on immediate posttests, with large effects.

In both of these studies, PSTM/WM were regressed on learning gains from very early exposures to new words. However, none of the comparative studies examining processing of new words (see Table 1) have either accounted for or utilized individual difference measures in PSTM/WM as predictors for processing or learning outcomes for new words. Given the focus of the experimental task in the present study to be on comprehension, and duration of exposure controlled through timed presentation of trials, I thought it both logical and necessary to measure and account for PSTM. Additionally, with the connection between multimodality in the treatment and outcome test items, it was considered desirable to include an auditory measure of PSTM, to examine the construct in a multidimensional way and account for these individual differences in processing and learning outcome models. However, given the variation for the present study in treatment conditions from intentional and incidental designs in other studies that found effects of PSTM, directional hypotheses and predictions were tentative. At minimum, it was important to account for PSTM as a potential predictor.

To date, given the variety and extent of research into input manipulations and their effects on vocabulary learning under intentional and incidental conditions, this work has been severely hampered by a near-total absence of accounting for individual difference factors in multidimensional constructs of proficiency and PSTM. As such, with research designs that focus on differences in cognitive processing behavior as well as learning outcomes, it is essential to include direct measurement of multiple dimensions both of proficiency and PSTM as potential covariate predictors of processing and learning. Since these constructs are both hypothesized to

have a profound influence on SLA (e.g., Skehan, 1991; Dörnyei, 2006), I considered it essential to include as much information on cognitive individual differences among participants in the present study. In this way, accounting for them would allow for more robust comparisons in RWL and RO behavior and learning, potentially clearing a path to more generalizable claims regarding behavior and outcomes among advanced-proficiency learners.

2.10 Overall Summary of Literature Review

In this literature review, I have provided the context for the present dissertation study. I summarized empirical findings to date regarding the processing and learning of novel L2 vocabulary under incidental conditions, and provided the theoretical underpinnings for how L2 lexical representations can develop during meaning-focused tasks. Additionally, I discussed the theoretical basis for examining the relative benefits of multimodal input as a form of unobtrusive input enhancement, the problematic and unidimensional way learning has been assessed to this point through visual-only test items, and how the psycholinguistic account for these benefits can and should be tested in empirical studies utilizing multimodal conditions. In short, very few studies in L2 contexts have gone beyond simply comparing learning outcomes in visual test items from single and multimodal conditions to make vague pedagogical recommendations. Even fewer have included specific and testable predictions regarding mechanisms for the reported benefits of multimodal input for L2 learners in meaning-focused contexts, while accounting for multidimensional individual difference constructs. The present study was designed to do both.

2.11 The Present Study

In the present dissertation study, I made and tested predictions based on research questions regarding both processing and learning outcomes for L2 readers encountering and learning novel pseudowords from RO and RWL during a meaning-focused, naturalistic, continuous story. I operationalized the incidental conditions by controlling for time on task, emphasizing the focus of the task on comprehension through providing comprehension questions prior to each chapter for preview, after each chapter to answer, and unannounced vocabulary learning posttests. I employed real-time, online measurement of processing through tracking eye movements to examine the trajectory of form-form familiarity across encounters with new words, and predicted differences between RO and RWL in that trajectory for the initial familiarity check (measured through GD), sentence-level meaning integration (TRT), and variability in encounters (visit count). Additionally, eye-tracking data helped to uncover the relationship between attention and learning gains, as well as differences by group (RO vs. RWL).

Additionally, I utilized both visual and auditory test items on three learning outcomes, following the theoretical sequence of form-form to form-meaning mapping during learning and lexical development, so as not to obscure the detection of potentially-emerging phonological form-form mapping benefits during RWL. Given the connection between reading ahead of the audio in RWL and L2 proficiency (Conklin et al., 2020), I made an initial prediction that reading ahead would result in duplicative effects of repetition on total reading time and learning gains. Finally, I included multiple measures both of L2 proficiency and linguistic/non-linguistic measures of PSTM to account for a wider range of cognitive individual differences than other similar studies have been able to include to this point.

2.12 Research Questions

The present dissertation study was designed to answer five primary research questions:

RQ1a: Is the online processing trajectory of form acquisition (as measured by eye tracking) for novel L2 words when reading different under incidental conditions during reading while listening, compared with reading alone?

RQ1b: To what extent is learning of novel word form and meaning different under incidental conditions during reading while listening, compared with reading alone?

RQ2a: To what extent does reading slightly ahead of the text in a reading while listening task contribute to learning of novel word form under incidental conditions, as evidenced by faster reading times?

RQ2b: To what extent does reading slightly ahead of the text in a reading while listening task contribute to learning of novel word form and meaning under incidental conditions, as evidenced by posttest outcomes?

RQ3: Accounting for individual differences in L2 proficiency, to what extent do individual differences in working memory influence vocabulary learning outcomes from reading while listening, compared with reading only?

2.13 Initial Predictions

When the present dissertation study was designed, there were no published studies that made comparisons of online processing behavior between RO and RWL groups. Since that time, Tuzcu (2023) was published, providing helpful comparative information regarding differences between processing in RO and RWL groups under incidental conditions, and allowing for more specific interpretations. Initial predictions for the present study were made without Tuzcu's (2023) findings, and were as follows (summarized in Table 2 below):

RQ1a: Is the online processing trajectory of form acquisition (as measured by eye tracking) for novel L2 words when reading different under incidental conditions during RWL, compared with RO, for L2 learners at high-intermediate to low-advanced L2 proficiency?

Prediction 1: Analyses of eye-tracking data will reveal significant differences in processing trajectory for novel word form-form connection learning, with participants in the RWL group demonstrating a greater decrease in GD and TRT across target word instances in a more linear and stable manner than in the RO group. It was expected that both groups would demonstrate reading times across exposures in an S-shaped developmental curve (see Godfroid et al., 2018; Tuzcu, 2023), indicating faster development of form familiarity from RWL. However, a main effect of group, and an interaction effect of group by instance was also expected, with participants in the RWL group taking less overall time to read the new words, due to less processing variability and greater reading fluency. Data visualization was expected to indicate divergence in GD and TRT particularly during later exposures to the novel pseudowords.

Prediction 2: Model analyses will reveal significant differences in visit count to target item interest areas, with participants in the RO group having significantly more visits to the targets than the RWL group. This would reflect greater variability in encounters with the novel pseudowords during RO, and an overall smoother reading experience in RWL.

RQ1b: To what extent is offline learning of novel word form and meaning different under incidental conditions during reading while listening, compared with reading alone?

Prediction 3: Mixed-effects models will reveal a main effect of group for form recognition posttest scores, with participants in the RWL group scoring significantly

higher than those in the RO group across each of the three outcomes. Additionally, a group by item modality interaction effect will indicate that learning gains are substantially higher for the form recognition outcome in the auditory modality, given that phonological information is simultaneously encoded with orthographic information during new word processing, and providing both in the RWL group will result in mutually strengthening form-form links in the newly-developing lexical representation (Bordag et al., 2022). As a brief sidenote, it was expected that learning of form would be evidenced in the auditory items for participants in the RO group, given participant proficiency and the importance of sublexical phonology when reading new words in English (see Brysbaert, 2022); thus, some form familiarity at the level of phonological encoding was predicted to be detected even among participants who were only exposed to the new words in the visual modality, especially given that the population for the study was at a high level of proficiency. However, form recognition learning outcomes were predicted to be significantly better in RWL than RO for auditory test items.

Prediction 4: Mixed-effects models will reveal a main effect of group for meaning recognition posttest scores, with participants in the RWL group scoring significantly higher than those in the RO group. Additionally, a group by item modality interaction effect will indicate that outcome scores will be higher in the RO group for visual versus auditory items, while no differences in item modality will be detected in RWL. This prediction was based on the assumption that phonological information is simultaneously encoded with orthographic and semantic information in RWL, resulting in mutual strengthening of form-meaning links in the newly-developing lexical representation

(Bordag et al., 2022) and a combination of phonological, orthographic, and semantic “bootstrapping” (Gor et al., 2021) in accurate word meaning identification.

Prediction 5: Mixed-effects models will reveal a main effect of group on meaning recall posttest scores, with participants in the RWL group scoring significantly higher than those in the RO group. Additionally, a group by item modality interaction effect will indicate that learning will be greater on visual items than auditory items for the RO group, but learning will be equal on both auditory and visual items in the RWL group. This prediction was also based on theoretical assumptions that phonological information is simultaneously encoded with orthographic and semantic information in RWL, resulting in mutual strengthening of form-meaning links in the newly-developing lexical representation (Bordag et al., 2022) and a combination of phonological, orthographic, and semantic “bootstrapping” (Gor et al., 2021) in accurate meaning recall. This prediction was made tentatively in the initial proposal, given the depth of knowledge necessary for L2-L1 translation (see Waring & Takaki, 2003), but the high level of proficiency for participants in the full study provided additional confidence that differences could be detected in the more difficult meaning recall outcome measure.

RQ2a: To what extent does reading slightly ahead of the text in a reading while listening task contribute to learning of novel word form under incidental conditions, as evidenced by faster reading times?

Prediction 6: Mixed-effects models will reveal a main effect of proportion of reading ahead as a significant negative predictor of total reading time across exposures, even when accounting for differences in proficiency and memory. This prediction was based on the hypothesized connection in the literature between form familiarity and a reduction

in processing time as new words become more familiar during the first 10 instances of them in a text, and the relationship between reading ahead of audio and L2 proficiency reported in Conklin et al. (2020). Since higher-proficiency and faster readers read slightly ahead of the audio more, it was hypothesized that when the overall proficiency level of participants is high (across the participant sample), across the time course of 10 instances within the text, initial encounters would bring more reading time from reading slightly ahead, but that RWL would facilitate faster form-form connections, and by extension faster subsequent TRT across all exposures, while controlling for reading speed. Since I controlled for reading speed as part of the multidimensional proficiency variable, any complexity in the nature of the relationship between individual differences in reading speed and a reduction in TRT over time based on reading ahead would be accounted for in analytical models. Given that no prior published study had made this prediction or tested it as a potential indicator of the psycholinguistic source of the benefits of RWL at the time of the dissertation proposal, it was exploratory in nature.

RQ2b: To what extent does reading slightly ahead of the text in a reading while listening task contribute to learning of novel word form and meaning under incidental conditions, as evidenced by posttest outcomes?

Prediction 7: Mixed-effects models will reveal a main effect of proportion of reading ahead as a significant predictor of form recognition outcome scores from items in both testing modalities within the RWL group, even when individual differences in proficiency and memory are accounted for.

Prediction 8: Mixed-effects models will reveal a main effect of proportion of reading ahead as a significant predictor of meaning recognition outcome scores from items in

both testing modalities within the RWL group, even when individual differences in proficiency and memory are accounted for.

Prediction 9: Mixed-effects models will reveal a main effect of proportion of reading ahead as a significant predictor of meaning recall outcome scores from items in both testing modalities within the RWL group, even when individual differences in proficiency and memory are accounted for. It was predicted that if participants were reading slightly ahead of the audio, initial form-form connections would be made during early exposures, and result in better form recognition scores, along with deeper semantic processing during subsequent encounters, resulting in greater meaning recognition and recall scores. Therefore, predictions 7-9 were exploratory in nature to examine a potential underlying mechanism for reported learning benefits of RWL.

RQ3: Accounting for individual differences in L2 proficiency, to what extent do individual differences in working memory influence vocabulary learning outcomes from reading while listening, compared with reading only?

Prediction 10: Mixed-effects models will reveal a group by PSTM interaction in both form and meaning recognition learning outcomes, with a resulting aptitude by treatment effect for both form and meaning recognition. This prediction was based on reported findings from other vocabulary learning studies that found ATI effects for PSTM on vocabulary learning under intentional (Martin & Ellis, 2012) and incidental (Malone, 2018) conditions. However, in previous findings, these effects were detected at very early stages of form acquisition, so this prediction was made tentatively.

Table 2*Initial RQs, Predictions, Tasks, and Measures in the Present Study*

RQ(s)	Prediction(s)	Group	Task/Measure	
			<u>Online Eye tracking</u>	<u>Offline measure</u>
1a	1-2	Both (RO and RWL)	RO > RWL in gaze duration and TRT, evidenced by significant group differences in GCA analyses RO > RWL in total visits to target interest areas	
1b	3-5	Both (RO and RWL)		RWL > RO on all posttest outcomes; significant group x item modality interactions for each of the three outcome measures (visual > auditory items in RO across all three learning outcomes)
2a	6	RWL only	Proportion of reading ahead as a significant negative predictor of gaze duration and TRT in RWL group	
2b	7-9	RWL only		Proportion of reading ahead as a significant predictor of all three posttest outcomes
3	10	Both (RO and RWL)		PSTM as a significant predictor of form and meaning recognition outcomes, interacting with group; strongest relationships with MR auditory items, given their deeper level of processing

Chapter 3: Methods

3.1 Introduction to Methods

In this chapter, an overview of all methods utilized to answer the research questions is provided, based on *a priori* predictions (see Table 2) and utilizing various analytic methods. The design will be described in detail, including participant demographics, task and item development, dependent outcome and covariate measures, procedure, and analyses.

3.2 Overview of the Research Design

Table 3 summarizes the way all variables were operationalized for the study, including between and within-group analyses. To answer RQs 1a, 1b, and 3, I compared processing and learning measures for RO and RWL groups utilizing the manipulation of input modality, while examining effects of PSTM on learning outcomes. In the RO group, participants had access only to the visual text, while in the RWL group, participants simultaneously listened to the spoken version of the text as they read it. To answer RQs 2a and 2b, data from within the RWL group only were included in the analyses.

Table 3

Operationalization of All Variables in the Present Study

	Variables	Operationalization	Instruments	Type of Data
Cognitive Processes	Attention	Eye fixations on interest areas including target words while reading the treatment text	Three eye-tracking measures (GD, TRT, visit count)	Continuous
		Reports of awareness and opinions on the learning task	Debriefing survey interviews	Categorical

Vocabulary learning measures	Form recognition	Accuracy of form recognition	Form recognition test	Categorical
	Meaning recognition	Accuracy of meaning recognition	Meaning recognition test	Categorical
	Meaning recall	Accuracy of meaning recall	Meaning recall test	Categorical
Learner-related factors	L2 proficiency	General reading and vocabulary ability	Cloze test (Brown, 1980)	Continuous
		Auditory vocabulary size	Auditory version of LexTALE (Lemhöfer & Broersma, 2012)	Continuous
		Reading speed	Number of words read per minute on first-pass reading of first two trials of summary of treatment story	Continuous
	Phonological short-term memory (PSTM)	Auditory non-linguistic PSTM	Running memory span (RMS)	Continuous
		Visual linguistic PSTM	Nonword span (NWS)	Continuous
General processing speed	Rate of speed-up in RT to spatial RT task	Serial reaction time (SRT)	Continuous	

3.2.1 Participants

For the study, given planned two-factor generalized linear mixed-effects analyses for dichotomous learning outcomes, I estimated optimal sample sizes based on Judd et al.'s (2017)

recommendations. Planned analyses included participants nested within condition, target items crossed with condition, and the two random factors (participants and targets) crossed (see the “NCC” structure on p. 17.12 in Judd et al. (2017) for example architecture). Input was set to standardized, and all Variance Partitioning Coefficients (VPCs) set to default levels. The optimal effect size d was set to 0.7, based on Plonsky and Oswald’s (2014) estimates of medium effect sizes in SLA research. Given the relatively underpowered nature of psychological research reported recently in Brysbaert & Stevens (2018), optimal power was set to Cohen’s (1962) benchmark of 0.8 or above. The number of participant observations was set at 50 per condition (100 total), with the minimum number of target outcome items 10 (for analysis at the level of exposure). Within these parameters, estimated power was 0.822, within set expectations for optimal power. I considered this to be a helpful starting point in estimating power, and interpreted the analysis optimistically that this number of participants and items would be sufficient for detecting true differences in the participant sample. However, given the relatively small number of other studies that have examined similar issues, along with the very small n sizes reported in similar studies (see Table 1), there was little precedent within Applied Linguistics/SLA research on setting optimal power. Therefore, power analysis for the present study should be considered exploratory in nature.

This estimated sample size is much larger than those for other similar studies (see Table 1). However, each of these studies reported track loss and other factors that resulted in the loss of participant data for the final analyses. It was expected that there would be track loss from participant eye-tracking data. Fortunately, the loss of participant data in the present study was minimal, with only one participant’s data in the RWL group excluded because of technical difficulties with the eye-tracking computer, which resulted in the participant having to repeat

multiple chapters of the story. None of the four studies mentioned above reported power analyses to determine the likelihood of detecting generalizable effects, although each reported and claimed effects between groups. The absence of *a priori* power procedures in ISLA vocabulary studies was recently noted as a significant limitation in the literature by Vitta et al. (2021); as the field of ISLA vocabulary research continues to grow, more informed power analyses can and should be utilized in study design (see also Nicklin & Vitta, 2021). Given the inevitability of practical constraints of the participant population, available funding, and time in SLA research (see Loewen & Hui, 2021), there is a tension in balancing sample size and power to detect differences.

For the participant sample, we collected full datasets from a total of 120 participants (60 in RO; 60 in RWL), among adult L2 English speakers studying or living near the University of Maryland, College Park. However, one participant's data session was interrupted near the conclusion of the task, and that participant's data were discarded prior to data analyses, so the reported participant demographics and dataset for analyses includes a total of 119 participants. Participant recruitment proceeded primarily through word of mouth on the UMD campus, with advertisements posted on listservs for graduate students, social media, through email lists for departments on campus with large numbers of multilingual English learners, and in English skill-based courses on campus. Institutional Research Board approval was received from the university for the research protocol, and participants were paid \$15 for completing Part 1 of the study (proficiency, memory, and language background), and an additional \$35 upon completion of Part 2 (story reading treatment, vocabulary learning outcomes, and debriefing survey).

Given the necessity of in-person meetings for the part 2 eye-tracking task, the population was unfortunately limited. Given the nature of the treatment task, I limited participation to those

who scored at least at a high-intermediate range on the visual cloze (29 or above on a 50-item test) and auditory LexTALE proficiency outcomes (37.5% or higher). Pilot data from participants who were enrolled in an intensive English program course at high-intermediate to low advanced levels of English had indicated that this level of ability on the proficiency measures was easily sufficient for text comprehension, and participants in the study scored very highly on reading comprehension questions from each chapter in the text, indicating that they were easily proficient enough to read the story (see participant descriptives below in Table 4). All participants recruited for Part 1 reached required proficiency benchmarks, and were invited back for the Part 2 experimental tasks.

Specifying predictions related to participant L1 would be a highly fruitful vein to pursue (see Discussion), but given constraints on the sampling pool, I did not limit L1 groups. This variability was accounted for by pseudo-randomized group assignment, with nearly identical L1 profiles for both RO and RWL treatment groups. Table 4 provides descriptive demographic data for the participants overall, while Table 5 provides L1 backgrounds for the participants. Five participants in each group had become fluent in either reading, speaking, or both in English prior to the age of 12, but over 90% of total participants became fluent in reading and/or speaking English at or after the age of 12. Proficiency scores were scrutinized for each of the 10 participants who became fluent at younger ages, and were determined to fit within the demographics of the entire dataset, so they were kept. Each participant who had become fluent prior to age 12 had been required to submit standardized English test scores for entry into a U.S. university for a degree program, which was considered additional evidence of comparable English background and ability with fellow study participants, even given the variation of physical age of fluency.

In terms of language background, there was remarkable diversity, with 22 unique L1s represented. English was the third or fourth language for 16 of the participants in the RO group, and seven in the RWL group. This can largely be attributed to the significant subset of the participant demographic in both groups born in highly multilingual environments in southern and eastern Asian nations. Again, given the comparable levels of English proficiency, these multilingual English learners were considered sufficiently similar both to participate in the study and have their data collected. However, as discussed below, the language background and proficiency levels of the sample for the full study were slightly different than those of the pilot study. All participants were pseudo-randomly assigned to one of two reading groups, RO and RWL, and participants from Bangladesh/India/Sri Lanka ($n = 34$ in RO; $n = 34$ in RWL) and Mandarin-L1 participants from China/Taiwan ($n = 19$ in RO; $n = 19$ in RWL) were divided evenly, in order to balance similar L1 script types across both groups. All participants reported normal or corrected-to-normal sight and hearing.

Table 4

Demographic Information of Participants

Variable	Reading only				Reading While Listening				Difference
	Mean	SD	Min	Max	Mean	SD	Min	Max	<i>d</i>
Age (years)	24.65	2.69	21	37	26.07	4.11	19	41	.40
LOR (months)	11.98	18.51	1	120	18.56	24.6	1	96	.30
Formal education (years)	17.13	2.63	12	29	17.86	2.78	12	27	.27
Age of spoken fluency	17.08	4.85	8*	28	17.1	4.87	7*	29	.004
Age of reading fluency	14.75	4.28	7*	24	14.68	4.78	7*	29	.02

*Five participants in each group had become fluent in reading, listening, or both in English prior to the age of 12.

Table 5*Language Background of Participants*

	Reading only	Reading While Listening
L1 background	Azerbaijani: 1	Amharic: 1
	Bengali: 1	Bengali: 4
	Cantonese: 1	Cantonese: 1
	Gujarati: 5	French: 1
	Hindi: 8	Gujarati: 3
	Indonesian: 1	Hindi: 10
	Kannada: 2	Indonesian: 1
	Kodava: 1	Malayalam: 5
	Korean: 2	Mandarin: 19
	Malayalam: 2	Marathi: 5
	Mandarin: 19	Portuguese: 1
	Marathi: 4	Sinhala: 1
	Portuguese: 1	Spanish: 1
	Tamil: 2	Tamil: 2
	Telegu: 9	Telegu: 3
	Yoruba: 1	Urdu: 1

3.2.2 Experimental task

The experimental reading task was a modified nine-chapter short story roughly 7,400 words in length, *How Much Land Does A Man Need?* by Leo Tolstoy (see Appendix A for the full task, including target pseudoword items). Modification of the text proceeded in the following steps:

- (1) Repeatedly-occurring individual nouns were identified as possible target replacements based on an analysis of the whole text using the Compleat web vocabulary profiler (Cobb, 2022). This analysis produced a list of 25 words or combinations of similar concrete nouns (animals, grains, money, measurements) that were repeated at least 10 times and could be replaced in the text by pseudoword substitutes. Nouns were selected given their importance to communicating meaning; comparisons with other parts of speech (e.g., Tuzcu, 2023) would be interesting and revealing, but given the complexity

of analytical models and variability at the level of participant, consistency in the target part of speech was considered desirable.

- (2) Once these words were identified in the text, their syntactic context was examined to ensure that they were occurring only within object positions within a sentence (direct object, indirect object, object of preposition). Research with English L1 readers has indicated that among nouns, items in subject position are more syntactically prominent and perceptually salient than those in object position (Moravcsik & Healy, 1998). Additionally, other studies (e.g., McKoon et al., 1993; Birch & Rayner, 2010) have reported that item and concept nouns in direct object (DO) position are more conducive to form encoding than nouns in object of preposition (OP) or indirect object position (IDO). Once substitutions and text manipulation of syntactic structure was complete, I examined the proportion of each syntactic context. In the final text, 41.2% of targets were in OP position, 15.8% were in IDO position, and 44% were in DO position.
- (3) After the targets were identified and modified into the pseudoword replacements (process described below), all other words (other than proper nouns) in the translated text above the most frequent 4,000 word families in English according to the BNC-COCA 25k word family list (Nation, 2017) were replaced with more frequent synonyms. This meant that all words other than the target items were well below the 8,000-9,000 range indicative of academic reading proficiency in English (e.g., Nation, 2006), and increased the likelihood that participants would know the other words in the text at the 96% or above range. This is the reading threshold which Waring and Takaki (2003) argued was necessary for adequate comprehension. Pilot testing and final study comprehension scores ($M > 95\%$) revealed that participants at this level could adequately follow the text and answer the

comprehension questions very accurately. Ultimately, 250 of the 7375 total tokens (items) in the text (not including proper nouns) were outside the most frequent 4,000 word families in English, meaning 96.6% text coverage within the first 4,000 word families, and each of these 250 tokens consisted of the pseudoword targets.

Once the text was established and modified, comprehension questions were created. For the final study, questions that resulted in accurate responses by all pilot participants were kept – three per chapter, and 27 questions in total. Appendix B includes the list of comprehension questions for the final study. The questions were previewed for each chapter before the participant read it, in order to direct focus to the meaning of the text, and participants keyed in a multiple-choice response to each of the three items for that chapter after reading it. None of the target items were included in or necessary to answer the comprehension questions.

3.2.3 Target item selection and pseudoword characteristics

Appendix C specifies lexical and contextual characteristics of the words which the target pseudowords replaced, along with the original sources (similar learning studies) utilizing the pseudowords. Concreteness ratings for each of the original words in the text were taken from Brysbaert et al. (2014), and were consistently very high (average mean = 4.67; SD = 0.3; min = 3.63, max = 5 on a 1-5 scale). Brysbaert and New's (2009) SUBTLEX-US frequency count data were used to examine word frequency for the original words, which varied substantially more than concreteness. Since these words would be replaced in the text, frequency was less important in consistency than concreteness, since concreteness involves the ability to infer information about a perceptible entity from a given context (e.g., Paivio & Csapo, 1973; Paivio, 2013), Pseudoword targets to replace these original words were taken from three similar studies on vocabulary learning (Elgort & Warren, 2014; Pellicer-Sánchez, 2016; Elgort, 2017), and were

orthographically and phonologically similar to English words. They were created by means of replacing only a single letter from an existing English word.

Tools from the English Lexicon Project (Balota et al., 2007) were used to examine lexical characteristics of the pseudowords, which can also be seen in Appendix C. They ranged from three to six phonemes, five to six letters, were one or two syllables in length, and had one to six orthographic neighbors. Spelling consistency was quantified in an adapted way from Chee et al. (2020), by calculating rime scores for each syllable and averaging them for an estimate of spelling consistency on a scale from 0-1. Spelling consistency ranged from 0.3-1.0, but was generally high overall (Mean = 0.77). Since distribution of occurrence was not a controlled variable of interest, the distance between instances in the text was distributed equally as much as possible, with mean distances ranging from 500-700 words apart in the text by word.

The role of contextual information in semantic encoding of novel words has long been hypothesized to be strong, both in L1 and L2 reading contexts (Dempster, 1987; Nagy, 1995; Schmitt, 2000). To examine contextual information in the story, I followed Elgort and Warren (2014) in designing a cloze measure of the entire connected text. I asked 12 native speakers of English to read through the story, filling in each blank where one of the pseudoword replacement targets would be with two logical real words, and giving confidence ratings for each guess on a scale of 1-6. High confidence ratings would indicate that the context for each would provide more information for supporting the development of form-meaning connections. Cloze answers were considered acceptable if they were in the correct category (e.g., for *Bancel*, originally *Tradesman* in the text, answers such as man / manager / businessman / director / VP were all considered acceptable). Of the 250 individual instances of the targets in the text, 21 (8.4%) did not achieve the 75% acceptability threshold. The semantic contexts for each of these 21 instances

were modified to provide additional contextual clues to the word's meaning, and a subset of six native speakers re-completed the cloze measure for those 21 instances. Round 2 accuracy rates were very high, typically 100%, and well above the 75% acceptability threshold.

Once the semantic contexts were established, confidence ratings on a scale of 1-6 for each correctly-identified word within each context were recorded. Confidence rating scores for all of the 25 targets were high across all 10 exposures, with the grand mean of confidence rating mean scores by instance at 4.63 (SD=1.28; min=3.88; max=5.08). Given the fact that participants in the study all encountered each of the pseudoword targets in exactly the same text/context, with comparisons between groups only occurring between RO and RWL, natural variation at the text and item level was considered acceptable and even preferable, given the attempt in this study to maintain ecological validity in reading a long, connected text with a variety of contexts for inferring meaning (Godfroid et al., 2018; Hamrick & Pandža, 2020). That being said, the confidence ratings indicated that the contextual information for each word provided substantial support for the development of form-meaning connections across instances with the targets.

Figure 1 indicates the visual pattern of confidence rating trendlines for each word by exposure. While there was some variability, the amount of contextual information was consistently judged to be high for semantic information, allowing participants to guess the meaning of new words from their context with a reasonable level of accuracy. Since all participants received the same text and context, these confidence ratings were not used as a predictor in analytic models, but it was important to establish that participants could guess something about the meaning for each word in its differing contexts, without a sharp increase or decrease in the amount of contextual information across instances within the trials. No dramatic

pattern of change was observed for any individual word across the ten instances, indicating a consistently high degree of confidence in contextual information.

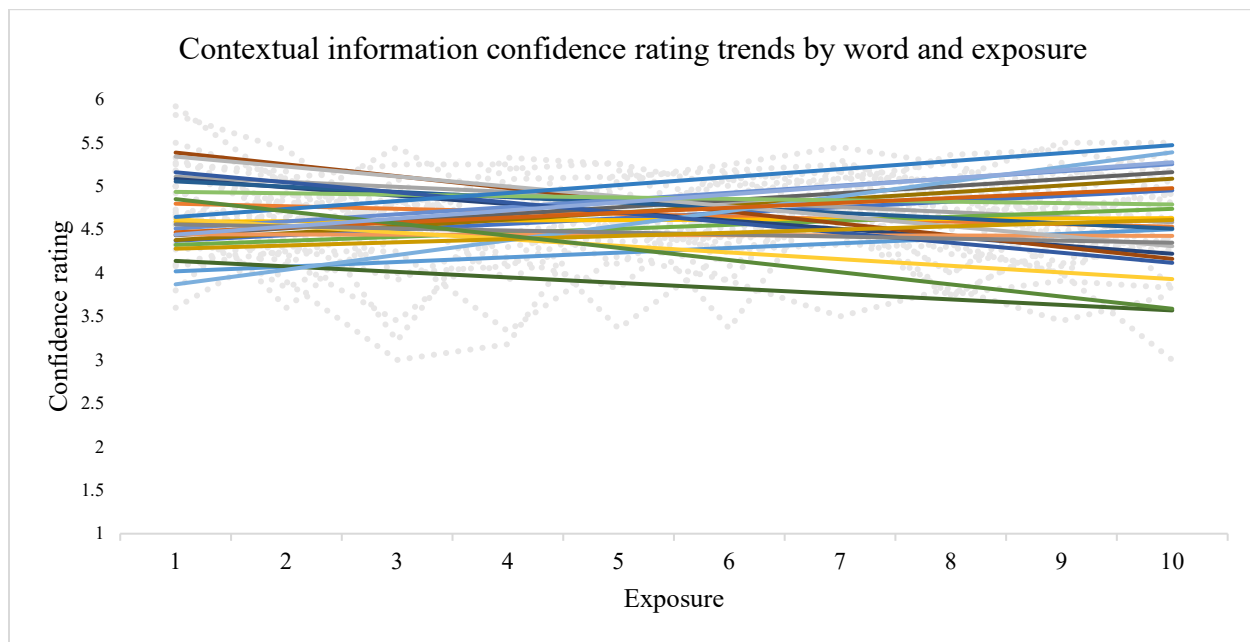


Figure 1. Contextual information mean rating trendlines for each target word context by exposure.

3.2.4 Vocabulary posttests

For the explicit vocabulary outcomes, posttest measures of both form-form and form-meaning connections were necessary, in order to examine initial stages of form recognition, as well as connections between form acquisition and semantic integration (for a discussion, see Webb, 2007; Webb & Nation, 2017). As such, three immediate posttests were created and piloted utilizing the Psychopy3 Builder interface (Peirce et al., 2019), and deployed via the online Pavlovia platform through a link from a Qualtrics survey. In each of the tasks, items were presented in randomized order and modality (visual or auditory). The three tasks were administered as follows:

- (1) Form recognition: 25 target pseudoword items and 25 pseudoword fillers (from the same studies from which the targets were taken) were presented one at a time, with the

question “was this word in the story?”. Participants responded by pressing a key for “yes”, or a different key for “no”. Participants were randomly assigned to either the “f” key for yes and the “j” key for no, or vice versa. Scores for this task were dichotomous, and the list of items can be found in Appendix D. For auditory items, participants were given the option of repeating the auditory item up to two times, in order not to disadvantage the auditory items compared with visual items, which had no time pressure.

(2) Meaning recall: the 25 target items were presented in randomized order and modality, and participants were provided with a paper list of numbered items with blanks. Participants were instructed to write an L1 translation of the target item in each blank, and write “I don’t know” if they had no translation even to guess. The complete list of items, along with the paper handout and screen instructions, can be found in Appendix E. As with the form recognition measure, participants had the option of clicking a button to repeat the auditory items up to two times. Scoring proceeded in two steps: (1) all written answers were translated using Google Translate for an initial score for each item by participant; (2) native speakers from the L1s represented in the study were asked to score a subset ($n = 58$) of the total participant written responses to the meaning recall task. Any items in which there was a discrepancy in translation resulted in the human rater’s scores being used. Inter-rater agreement between human raters and Google Translate was extremely high for the meaning of the translations (>0.99). Any differences between reported scores resulted in human rater scores replacing Google Translate scores, and the very high inter-rater agreement was considered evidence that the remainder of the translations from Google Translate were accurate. The meaning recall task was

completed second, in order not to advantage participants in meaning recall scores by having seen the correct word meaning during the meaning recognition task.

- (3) Meaning recognition: 25 target items were presented in randomized order and modality, with three multiple-choice possible answers for the meaning for each pseudoword item (the correct answer and two distractors). Meaning was defined as the semantic category of each word, and all items can be found in Appendix F. As with the other two measures, participants had the option of clicking a button to repeat the auditory items up to two times. Again, scoring for this task was dichotomous correct/incorrect.

Reliability estimates for each explicit vocabulary learning outcome were initially calculated using Rasch IRT analysis with the *snowIRT* package in R using the Jamovi GUI (Seol, 2020), in order to make item-level comparisons of the fit of the data to predictions, for the purpose of computing item and person-level reliability coefficients (see Table 10 below in Results for reliability estimates).

3.2.5 Individual difference measures

I utilized three English proficiency measures, two PSTM tasks, and a task of general processing ability as indicators of cognitive individual differences. For proficiency, the first two measures were coded into a Psychopy experiment (Peirce et al., 2019) prior to the study, and presented through its online Pavlovia platform. First, an untimed English cloze measure of reading, vocabulary, and grammatical ability (Brown, 1980) was completed (see Appendix G). This measure is visual in nature, and provides a global assessment of English reading and vocabulary ability. It is scored according to a specified and validated key, and participants were scored strictly according to acceptable answers in the key both for lexical and grammatical accuracy. The maximum score for the cloze measure was 50.

Secondly, participants completed an auditory version of the LexTALE task (Lemhöfer & Broersma, 2012; see Appendix H). This is a lexical decision task validated for participants at an advanced level of L2 English; it was initially developed and validated for the visual modality, but I converted it into a more difficult auditory lexical decision task utilizing the materials and design of the original. A female native speaker of English recorded all items (real and nonwords) twice, with one chosen as the better exemplar, and audio was standardized utilizing Audacity recording and editing software (Audacity team, 2023). The order of items on the auditory LexTALE task was unchanged from the visual version, given the specifications of the test, and consisted of 60 un-speeded trials. As with the vocabulary outcomes, participants were randomly assigned to a target key pair order (“f” and “j” for either “yes” or “no”), consistent for each participant across all items. Scores for LexTALE were computed by calculating the proportion of correctly-identified words (out of 40) and correctly-rejected nonwords (out of 20) correctly answered, multiplied separately by 100, and then added together. The score for the auditory LexTALE task was the total proportion of correct answers, then, is the weighted proportion of answers (on a 0-1.0 scale).

The final proficiency measure was a simple measure of reading speed. Participants were instructed to read through a brief, two-screen introduction to the short story (see Appendix I). Their first-pass reading times for each word on the screen were recorded for this brief passage, and used to determine reading speed as a continuous variable in words per minute. This task was visual only in nature, in order to standardize comparisons between participants in RO and RWL in their overall reading speed without the additional variable of simultaneous audio.

Two measures of PSTM were employed. First, an auditory running memory span task was utilized (Pollack et al., 1959; see Bunting et al., 2006). In it, participants heard a list of 12 to

20 Roman letters, at a rate of three letters per second. Participants were instructed to recall and enter the most recent six letters they heard, and accurate responses included both the correct letters and their ordered positions. Secondly, participants completed a visual nonword span (NWS) task of PSTM (Gathercole et al., 2001; Linck et al., 2013). In this task, participants saw seven consecutive nonwords (utilizing Roman letters) presented visually one at a time. Nonwords were then presented on the screen, and participants were instructed to key in a response of whether the nonword was in the most recent list. Given the access I had to both of these tasks, and the fact that they are presented in separate modalities, it was considered beneficial to include both.

Finally, I used a measure of general processing ability, the serial reaction time (SRT) task. I used the same measure employed in Bonilla et al. (2020), which was adapted from Willingham et al. (1989). In this task, four boxes arranged horizontally on the screen indicate four possible locations for an asterisk to appear, and each box corresponded to a specific key (Z, X, C, or V). A trial consisted of an asterisk appearing in a box, and the task was to press the corresponding key as quickly and accurately as possible. Once the correct key is pressed, the asterisk immediately moves to another random location of the three other boxes. There were six blocks of 96 trials each, and the score was the mean of median RTs of the first block, presented in randomized order, with lower scores indicating faster processing speed (following Bonilla et al., 2020 and Linck et al., 2013). Access to, scoring of, and reliability metrics from the two PSTM tasks and the SRT measure was graciously provided through the Applied Research Laboratory for Intelligence and Security (ARLIS) Measurement Platform at the University of Maryland (see Table 10 in Results for reliability estimates).

3.2.6 Procedure

Data collection proceeded in two parts. First, participants were recruited by word of mouth on and around the University of Maryland for the Part 1 in-person sessions to complete the English proficiency, memory/processing, and language background survey tasks. The planned sample for the study included a minimum of low-advanced levels of English ability, given the need for familiarity with the most frequent 4,000 word families in order for comprehension of the treatment text. Pilot participants all scored a minimum of 40% on the auditory LexTALE task, and a minimum of 50% on the cloze measure. Since these participants had no issues in understanding the story and answering the comprehension questions accurately during the pilot, the full study used these scores as an initial threshold for qualification for the Part 2 tasks. With the exception of one participant who scored at 37.5% in the auditory LexTALE task, these thresholds were kept, and all participants demonstrated strong reading comprehension scores during the Part 2 reading.

Table 6 summarizes the procedure, including all tasks, for Parts 1 and 2 of the study. The description of the study on the informed consent form for both Parts 1 and 2 was deliberately vague, due to the nature of the study and its requirement of strict incidental conditions. The language background questionnaire, based on LEAP-Q (Marian et al., 2007), revealed the age of spoken and written fluency for each participant in English. Given hypotheses regarding sensitive periods of language learning in age of acquisition studies (e.g., Granena & Long, 2013), it was desirable that participants for the study became fluent in speaking and reading in English after the age of 12. With the exception of five participants balanced across each group (see Table 4 above), most participants in the full study met these benchmarks. The language background questionnaire can be found in full in Appendix J. Each of the proficiency and memory tasks were

embedded through links from a Qualtrics survey, and participants were able to navigate it easily. For each participant, progress was monitored actively during the tasks, with a researcher available to answer questions and give clarification as necessary. Given the time frame for completion of these tasks during the pilot study, it was expected that participants would take 60-75 minutes to complete the Part 1 tasks in full, which was reasonably accurate for the full study. Participants were paid \$15 for completing Part 1 tasks. If Part 1 participants voluntarily chose not to participate in the part 2 tasks, their Part 1 data was discarded as additional pilot data for the measures.

Participants who completed Part 1 and were interested in completing Part 2 were invited to meet at the eye tracking lab for the program in Second Language Acquisition at the University of Maryland, College Park. In the lab, participants were pseudo-randomly assigned to one of the two treatment groups (RO or RWL) based on L1. Calibration and orientation with the eye-tracking device and the task proceeded individually with participants. Participants (1) previewed the story by reading a summary (with reading speed measured), (2) previewed reading comprehension questions prior to each chapter, (3) read or read while listening to each of the nine chapters in the story, and (4) answered the reading comprehension questions after each chapter. Task instructions at the beginning of each chapter reminded participants in the RWL group to read along with the audio, and participant eye movements in the RWL group were monitored to ensure that they were generally synchronized with the auditory recordings. If synchronization was noticeably off, participants were gently reminded by the attending researcher to read along with the audio.

Offline vocabulary learning posttests were unannounced, and given to participants on a separate computer immediately following completion of the reading task. A short debriefing

session followed the experimental tasks on the same Qualtrics survey, wherein I elicited qualitative feedback on the reading and learning experience with the new words, asked specific questions to the RWL group about the benefits and challenges of simultaneous audio, and explained the full purpose of the study to each participant (see Appendix K for the debriefing section questions). Pilot testing revealed that the entire Part 2 process would take approximately two hours to complete, and participants in the full study nearly always completed the tasks within two hours. Participants were paid an additional \$35 for completion of Part 2, for a total payment amount of \$50 for participants who completed both parts.

Table 6

Study Tasks and Procedure in Chronological Sequence

Part	Task(s)	Measure(s)
1	Informed consent	N/A
1	Proficiency tasks	Cloze measure; auditory LexTALE
1	Phonological short-term memory / general processing	RMS; NWS; SRT
1	Language background questionnaire / participant payment	N/A
End of Part 1; Part 2 scheduled with each participant on a subsequent day		
2	Informed consent	N/A
2	Reading task with comprehension questions (random assignment to groups)	Decoding speed (summary page); Comprehension questions
2	Vocabulary posttests	Form recognition; Meaning recognition; Form-meaning recall (both auditory and visual items)
2	Debriefing / clarification on study purpose / participant payment	N/A

3.2.7 Apparatus

Eye movements were recorded with an EyeLink 1000 desk-mounted eye-tracking camera, which sampled eye movements for each participant's right eye at 1000 Hz (SR Research Ltd.; <https://www.sr-research.com/>). During the reading task, participants sat in front of a computer screen (1920 x 1080, with a 60Hz refresh rate) at a distance of 66 cm. Head movement during the task was minimized through the use of a chin and forehead rest. Participants were given brief breaks from the head rest after each chapter of the story, and then eye movements were re-calibrated prior to each new chapter. Given the length of the story and the unique challenges of reading in this way, breaks were necessary for participants to remain focused on the task. However, since the content being read was a continuous story, participants became increasingly eager to learn what happened in the story as it built to its narrative climax. This allowed for a more ecologically valid assessment, given the continuity of reading a story that one might pursue when reading for pleasure or in a language class.

The story and comprehension questions were divided into 190 trial screens across the nine chapters, with all text (including comprehension questions and instructions) triple spaced and presented in 18-point Courier New font. The text on each screen was left-aligned with a one-inch margin inserted around the edges of the screen to minimize spatial recording errors (Godfroid, 2020b). Zero to four target pseudowords were included on each screen, depending on the distribution of the target pseudowords in the story, and none were presented less than four words apart from one another. To control for eye-tracking measurement errors, with the exception of one instance of the target item *lastor*, the target pseudowords were not presented in the first or last line, as the first or last word in a line, nor as the first or last word in a sentence (Rayner & Pollatsek, 2016; Godfroid et al., 2018; Godfroid, 2020b). For the one instance of one

word where the target was erroneously placed as the final word in the sentence, the eye-tracking data for that instance was discarded for all participants (119 total data points, or 0.04% of the total eye-tracking data). Each comprehension question was presented at a separate screen individually.

Since time was controlled in this study, each screen was visible for participants in both groups only for the duration of the auditory recording of the text on each screen, with a brief two-second pause after the audio concluded prior to automatic progression to the next page. Participants in the RO group were instructed that the text would only be presented for a short time, and they should read quickly and carefully to understand the meaning of the passage. No participant in RO had difficulty either completing the reading for each screen or correctly answering the comprehension questions. Each screen included 2-7 lines of text, presented in monospaced Courier New font for psycholinguistic consistency in letter width (see Conklin & Alotaibi, 2023). Auditory recordings of each page varied in length from 20-39 seconds per page, and prior to the presentation of each page, a brief drift check was conducted by participants fixating on the upper left-hand corner of the screen. Once the time for each screen was complete, the screen automatically moved to the next screen. At the conclusion of each chapter, participants progressed through individual screens answering the multiple-choice comprehension questions by pressing the correct answer key (a or b). Following the multiple-choice questions, the story screens paused until the participant had a short break to relax, and indicated they were ready to go on.

All sessions began with instruction screens, followed by a calibration screen, and a nine-point calibration was performed at the beginning of each reading session and after each break. Each screen began with a drift correction, which examines the extent of spatial recording error.

For drift correction, a black dot appeared at the top-left side of each new screen prior to the next trial, and participants were asked to look at the center of the dot. The text for the subsequent screen was presented after a successful drift correction. During the entire experiment, we monitored the data recording quality from the host computer, and performed additional calibrations whenever we detected any drifts of eye movements, either during drift correction or drifting from the text line while participants were reading. This was a very infrequent occurrence, which preserved a more continuous reading experience for participants.

3.2.8 Auditory recordings and speed

For the auditory recordings, a female native speaker of English recorded each chapter fully. The recordings were divided by screen and manipulated using Audacity recording and editing software (Audacity Team, 2023), in order to ensure that rate of speech fit with participants' ability to read concurrently without participants reading too far ahead or reading too slowly to catch up to the audio. For pilot testing, participants in the RWL group were engaged well when the recording rate ranged from 140-160 words per minute. This fit comfortably within normal parameters for auditory speech, below what Pawlas et al. (1996) labeled as L1 "rapid" speech rate above 160 words per minute; however, the audio seemed to lag at times during the pilot study for the RWL group, so the optimal rate of 160-180 words per minute per screen was set for the final study. These norms follow better with the slightly higher reported rates in Goldman-Eisler (1961) and Griffiths (1990), which Conklin et al. (2020) followed. Given the normal rate of reading in both L1 and L2 typically being faster among learners of similar proficiency levels (See Conklin et al., 2020), this rate of speech was designed to give participants in the RO group a comfortable amount of time to read each screen for meaning, while not encouraging participants in the RWL group to read too far ahead. Some of the potential

implications of these decisions are discussed below, especially in light of the higher-than-expected level of English proficiency, and how results from another recent study comparing RO and RWL group (Tuzcu, 2023) may suggest that eye movement data in multimodal conditions may be impacted both by task instructions and audio speed. Audio recordings were standardized with each other for volume, although participants were given a volume check for the computer system and asked if they could hear well before the beginning of the study and between chapters. All participants in the RWL group reported being able to hear the recording well.

3.3 Data Analysis

3.3.1 Eye-tracking data pre-processing and cleaning

The default cognitive configuration of Eyelink was used for the parsing of eye-movements in the main reading text (Godfroid & Hui, 2020). A technical issue occurred for one participant, resulting in a repetition of the first eight chapters of the story. As a result, this participant's data were removed from the datasets, both for eye tracking and learning outcomes. All eye-tracking data were cleaned using the default four-stage fixation cleaning procedure of the Eyelink DataViewer program (SR Research Ltd.; <https://www.sr-research.com/>). In the first two steps, overly short fixations were merged with neighboring fixations (i.e., fixations immediately before or after) within a specified distance based on a threshold value (first stage = fixations <80 ms and within 0.5 degrees; second stage = fixations <40 ms and within 1.25 degrees). Then, during the third step, any instances of three consecutive fixations <140 ms in an interest area were merged into one fixation. Finally, any fixations lower or higher than specified thresholds were removed from the data. After checking the removal threshold values used in previous studies and examining fixation durations in the current dataset, I set the parameters at fixations <80 ms and >800 ms from the dataset, as these fixation values reflect minor location errors rather

than cognitive processing and losses of concentration, respectively (Carrol & Conklin, 2020; Pellicer-Sánchez, 2016; Godfroid, 2020b). However, this resulted in no loss of data. I also visually inspected the remaining eye-tracking data trial-by-trial for any track loss or drift instances (Godfroid, 2020b). No track loss was detected, and any vertical drifts were corrected manually. The initial pool of eye-tracking data for the target pseudowords contained 29,750 data points (250 interest areas for target pseudowords \times 119 participants). After the data preprocessing and cleaning stages, 119 data points were discarded (0.04% of the total data), leaving 29,631 data points.

During model criticism of gaze duration (GD) and total reading time (TRT) data, skips of individual instances were also removed from the dataset, in order to minimize skew in GD and TRT. This resulted in an additional reduction of 855 data points (2.9% of the total eye tracking data), resulting in 28,776 data points for fixations in the target interest areas (pseudowords). The threshold of ± 2.5 standard deviations was set for the removal of outliers in the eye tracking dataset, which resulted in the removal of an additional 530 total data points, resulting in a grand total of 28,246 data points for the final analysis of GD and TRT.

3.3.2 Individual difference variables

Since the first four research questions involved accounting for individual differences in proficiency, memory, and processing speed, the first steps in model building for the analyses were to determine to what extent these variables correlated with each other. I followed Plonsky and Oswald's (2014) recommendations for correlation coefficient thresholds, with small/medium/large effects established at the threshold of $r_s = .25, .40,$ and $.60$ respectively. If the correlations were very large, it would be evidence of collinearity. If they only correlated moderately, it would be evidence that they were tapping into similar but distinct constructs.

Table 7 indicates the Pearson correlation coefficients between the proficiency variables (cloze, auditory LexTALE, reading speed), which were either very small to medium in size. Next, the Pearson coefficient for the relationship between the two PSTM measures (RMS and NWS) was 0.17, indicating a weak relationship. Given that these relationships were clear, but relatively weak, I initially created a composite variable for proficiency for each participant by averaging the three transformed z-scores of the three measures, to use as a single metric in the final models for the sake of parsimony. The same process was then conducted for the two PSTM measures, resulting in a composite PSTM variable for analytic models. For both composite variables, I did not have theoretical reasons to weight them differentially, so this approach assumes equal weighting from each sub-component of the variable (proficiency and PSTM), and that each measure taps into the same underlying construct (English proficiency and PSTM, respectively). Additional exploratory analyses were undertaken utilizing each individual difference measure as a distinct predictor, due to the lack of strong relationships between individual difference factors, in order to clarify whether their relative contribution could be determined (see below for these follow-up analyses for main effects by RQ).

Table 7

Pearson Correlation Coefficients between Proficiency, Memory, and Processing Variables

	Cloze	Auditory LexTALE	Reading speed (words per minute)	Running memory span (auditory PSTM)	Nonword span (visual PSTM)	Serial reaction time
Cloze	1	0.44	0.22	0.26	0.32	0.01
Auditory LexTALE	0.44	1	0.15	0.08	0.11	0.23

Reading speed (words per minute)	0.22	0.15	1	0.18	-0.06	0.02
Running memory span (auditory PSTM)	0.26	0.08	0.18	1	0.17	-0.24
Nonword span (visual PSTM)	0.32	0.11	-0.06	0.17	1	-0.25
Serial reaction time	0.01	0.23	0.02	-0.24	-0.25	1

There were three participants whose scores on the SRT task were missing from the dataset due to technical issues. In order to account for them, I utilized a random forest imputation algorithm for missing data, the *MissForest* algorithm (Stekhoven & Buhlmann, 2012) implemented by the *missForest* package in R. The algorithm was applied to the whole dataset for all included variables, imputed initial mean/mode information for missing data, and then utilized a random forest method to make predictions for representing ability in the missing data. That is, the algorithm does not make predictions about the form of the function, but instead estimates the function iteratively in a way that approximates the data points. This resulted in estimated scores for the three missing values, which were included in the total dataset for the three participants for the processing speed (SRT) measure.

Figure 2 indicates distributions and Pearson correlation coefficients between the composite proficiency and PSTM variables (averaged z-transformed scores for the three proficiency measures and three two PSTM measures, respectively), and z-transformed scores on the measure of general processing speed (SRT). Since these correlations were in the small-medium range, there was no evidence of collinearity, and these measures were utilized as the

covariate predictors in the models.

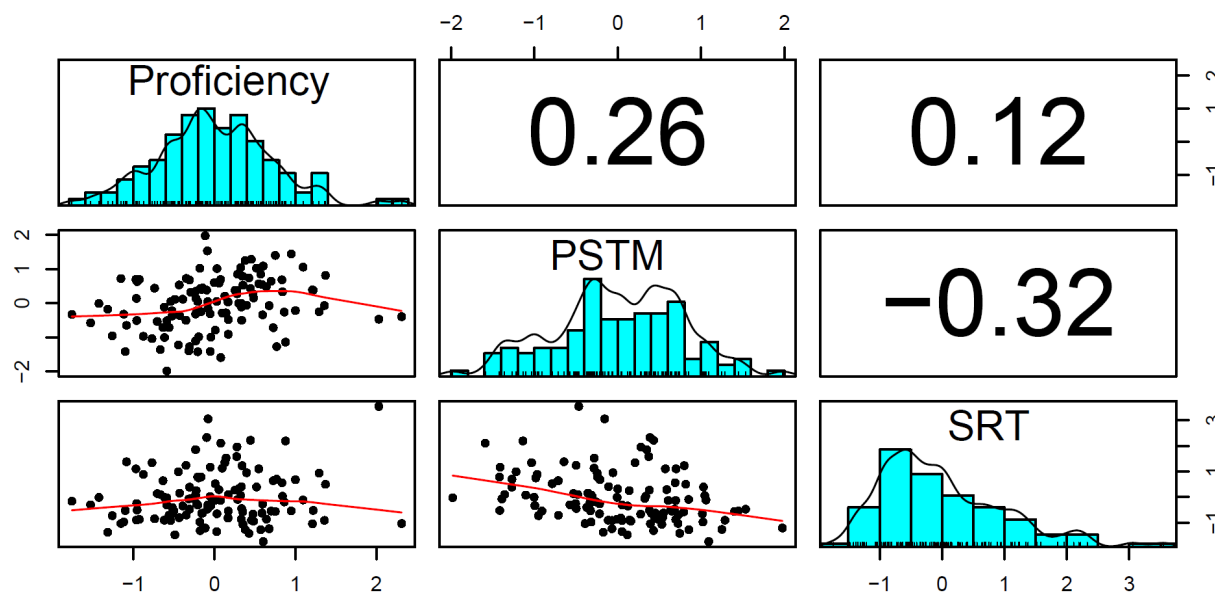


Figure 2. Correlation coefficients and graphs for the three individual difference covariates.

3.3.3 Eye-tracking measures as outcomes for between-group analyses

Three eye-tracking measures were utilized in order to compare group-level processing behavior on reading new words to answer RQ1a. First, gaze duration (GD), or the total time spent during the first pass through an interest area, was measured. This is considered an early measure of word familiarity (see Godfroid, 2020a), and has been utilized as an estimate of the familiarity check stage of word identification within the EZ-Reader paradigm, along with use as a time-course measure of new word learning (i.e., form familiarity) in recent L2 vocabulary learning studies examining processing (e.g., Pellicer-Sánchez, 2016; Elgort et al., 2018; Godfroid et al., 2018; Tuzcu, 2023). Secondly, I measured total reading time (TRT) as an indicator of total attention on the target items. Finally, in the proposal for the study, I indicated that I would utilize *regression count* data as an indicator of reading stability across time. However, regression count data simply include all cases where eye movements return directly to an interest area from a

subsequent interest area (regression *in*), or all cases where eye movements exit the target interest area and return to an earlier interest area (regression *out*). Taken together, these two measures are helpful, but do not include cases where the reader moves through the initial encounter with the word, and returns to the larger phrase or sentence prior to the target for broader comprehension, and prior to re-entering the target interest area. As I reflected on the different options for capturing *all* encounters with the new words, which L2 vocabulary learning is highly sensitive to under incidental conditions (e.g., Rott, 1999; Brown et al., 2008), a better option emerged. *Visit count* data include all cases of entering into the target interest area, either from prior or subsequent interest areas, and better capture the full number of encounters with a new word. As such, it was utilized in place of regression count data as the late eye-tracking count measure, and an indicator of reading stability and semantic integration processes (see Godfroid, 2020b).

Prior to conducting inferential analyses, eye-movement data for each group were inspected using (a) the plots of mean eye tracking measures (GD, TRT, and visit count), and (b) plots illustrating patterns in the data. This was done by applying smoothers utilizing the loess method, reflecting longitudinal effects of order of occurrence (from the first to the tenth instance of the target items in the text). Separate plots were created for each relevant eye-tracking measure – GD, TRT, and visit count (see Results below).

Initial plots of GD and TRT indicated possible non-linear (S-shaped) changes in participant eye movements on target pseudowords across 10 exposures. As a result, following the initial plan from my dissertation proposal and other recent similar work (Godfroid et al., 2018; Tuzcu, 2023), repeated occurrences were treated as a time-course variable to conduct growth curve analyses for GD and TRT (Mirman, 2014; Godfroid, 2020b). The primary goal of these analyses was to examine changes in GD and TRT over time across the ten exposures to the

pseudowords embedded in the text of the story. The non-linear decreasing trajectory of GD and TRT reported in other studies of L2 lexical development while reading (e.g., Pellicer-Sánchez, 2016; Godfroid et al., 2018; Elgort et al., 2018; Tuzcu, 2023) has been used as a measure of the development of form knowledge for new words, so the same basic structure was initially employed in this study's analyses. In growth curve models, the linear term of a variable (Instance¹ in this study) higher polynomial terms including quadratic (Instance²) and cubic (Instance³) were also included during initial model building as predictors.

Model fit for GD and TRT proceeded using the *lmer* function of the R package *lme4* (Bates et al., 2015). During model fitting, restricted maximum likelihood (REML) was used as the estimation method, with GD and TRT as the outcome variables. As the distributions of GD and TRT data are typically (and were in this study) positively skewed, they were log transformed with base e. Independent variables in the base omnibus models were as follows: group (dummy-coded RO vs. RWL), linear instance, quadratic instance, cubic instance, and the interaction term linear instance by group. Visualization of graphs for GD and TRT indicated a possible S-shaped pattern in the data, particularly for the RO group, so quadratic and cubic terms were initially included. All exposure terms were changed into orthogonal polynomials prior to adding them into the models, in order to decrease collinearity between instance terms in the GD and TRT analyses. However, as discussed below, the best-fitting model term according to chi-square tests of fit utilizing the *anova()* model function in R included only the linear term, so for the purposes of parsimony, the orthogonal terms were dropped, and the best-fitting models for GD and TRT were based on linear mixed-effects regression.

Model fitting for the visit count variable was based on given the tendency of count data to be non-normally distributed and over-dispersed (e.g., Lawless, 1987; Ver Hoef & Boveng,

2007), so model building for visit count data proceeded through a negative binomial generalized mixed-effects framework, utilizing the *glmer* function of the R package *lme4* (Bates et al., 2015). The fixed effect for the model was group, with covariate terms proficiency, PSTM, and SRT included. The interaction term group by instance was included in model building, as well, to indicate whether differences in visit count varied across instances by group (RO vs. RWL). All intercept and slope variance components were specified, and included if they contributed to overall model fit.

Continuous covariate participant-level variables proficiency, PSTM, and processing speed (SRT) were entered as fixed effects to all ET data models. Additionally, all intercept and slope variance components were specified both for ET and visit count modeling, and included if they contributed to overall model fit. For all eye tracking data models, random slopes were tested, but did not contribute to model fit (either through non-convergence or chi square tests), and are not presented in the final models below. Effect sizes for the group comparisons were reported as Odds Ratios for visit count and Cohen's *d* for all eye tracking measures.

3.3.4 Vocabulary learning outcomes for between-group analyses

The vocabulary learning outcome data were dichotomous in nature, scored 1 (correct) or 0 (incorrect). To analyze learning gains on each of the three outcomes (form recognition, meaning recognition, meaning recall), I built logistic mixed-effects regression models (Jaeger, 2008) using the *glmer* function in the *lme4* package in R (Bates et al., 2015). Scores on the distractor items were very high for all participants on the form recognition task (70-85% accuracy), so main analysis for the form recognition task consisted of the dichotomous scores on target items only, following two similar recent studies (Godfroid et al., 2018; Tuzcu, 2023).

For all vocabulary learning outcome models, group (RO vs. RWL) and item modality (visual vs. auditory) and their interaction were main categorical predictors in the models, with proficiency, PSTM, and processing speed (SRT) as covariate predictors in maximal models. Comparability between individual difference predictors was achieved through z-transformations to center and scale them. Random intercepts at the participant and item level were included, as well. For all vocabulary learning outcome models, random slopes were tested in the initial maximal models, but either did not contribute to model fit or resulted in model non-convergence, and are not presented in the final models below. Model selection proceeded in a backward, stepwise selection of predictors and interactions. Model comparison was made from chi square tests of fit in the mixed-effects models, with best-fitting models being the most parsimonious with the fewest covariates. I used the odds ratios provided by the outcome models to standardize and report Cohen's *d* alongside odds ratio metrics for each coefficient estimate for effect size.

3.3.5 Reading ahead of the audio as a predictor of TRT and VL in RWL

The reading-ahead variable was operationalized within the RWL group data as the proportion of looking ahead of the audio at the time of auditory presentation of the new words in context, summed across all individual instances with each word. This created a continuous main predictor of proportion of reading ahead (readahead) for model building, and I specified only the RWL dataset for these analyses. First, I built a linear mixed-effects regression model, utilizing TRT as the outcome. Secondly, I built three separate logistic mixed-effects models with the vocabulary learning test scores as the outcomes. For all readeahead analyses, I again utilized the *lmer* function in the *lme4* package in R (Bates et al., 2015), with the proportion of reading ahead as the main predictor, and the individual difference variables as the covariates. Model building proceeded in an equivalent manner to the group-level outcome comparisons, using backward

stepwise model selection, with the best-fitting models being most parsimonious with the fewest predictor variables. For the linear mixed-effects model with TRT data as the outcome, effect sizes were reported as Cohen's *d*; both odds ratios and Cohen's *d* were reported below as effect sizes for the logistic models. Again, random slopes were tested in the initial maximal models for the readahead variable analyses, but did not contribute to model fit (either through chi square tests of fit or non-convergence), and are not presented in the final models below.

3.3.6 Qualitative survey data coding

Even though the debriefing survey was not related to a specific research question, the committee for the proposal indicated that expanding the questions on it would provide interesting (albeit informal) insights into how aware participants were of the target items. Given the targeted focus of questions on the debriefing survey data (see Appendix K), I utilized basic qualitative coding techniques for these data, along the lines of the first stage of Strauss and Corbin's (1997) grounded theory, *open coding*. This method involves creating concept labels and categories for descriptive data, involving significant thematic elements. Given the relative brevity of questions and responses in the qualitative dataset, this was considered sufficient for examining emerging themes from participant perceptions on the story, encounters with novel words, and perceptions from the RWL group about the relative benefits of the audio on reading and learning new words.

3.4 Summary of Methods

Table 8 summarizes all statistical models utilized to answer the research questions, including relevant variables and syntax for maximal models for the quantitative data analyses. In this chapter, I have summarized the methods employed for the dissertation study. I described the participants, experimental reading task, target pseudoword item characteristics, vocabulary

learning posttests, covariate individual difference measures, procedure, eye-tracking apparatus, and methods for recording audio. I then described the planned statistical analyses for each research question/focus area, as well as the plan for examining qualitative data. The results of these analyses will now be presented.

Table 8

Statistical Models Utilized in the Study by RQ and Prediction

RQ	Prediction(s)	Model	Measures / effects	Syntax for maximal model
1a	1	Growth curve analysis	<p>Outcome: gaze duration Main effect: group (RO vs. RWL) Interaction term: group x instance Covariate fixed effects: proficiency, PSTM, SRT, Visit Count Random effects: person, item</p>	$GD \sim 1 +$ $Group * Instance +$ $Proficiency + PSTM +$ $SRT + Visit\ Count + (1 $ $Person) + (1 Item) +$ $(Instance Person) +$ $(Visit\ Count Person) +$ $(Group Item) +$ $(Instance Item) +$ $(Proficiency Item) +$ $(PSTM Item) + (SRT $ $Item) + (Visit\ Count $ $Item) + (Group:Instance$ $ Item)$
1a	1	Growth curve analysis	<p>Outcome: total reading time Main effect: group (RO vs. RWL) Interaction term: group x instance Covariate fixed effects: proficiency, PSTM, SRT, Visit Count Random effects: person, item</p>	$TRT \sim 1 +$ $Group * Instance +$ $Proficiency + PSTM +$ $SRT + Visit\ Count + (1 $ $Person) + (1 Item) +$ $(Instance Person) +$ $(Visit\ Count Person) +$ $(Group Item) +$ $(Instance Item) +$ $(Proficiency Item) +$ $(PSTM Item) + (SRT $ $Item) + (Visit\ Count $ $Item) + (Group:Instance$ $ Item)$

1a

2

Negative
binomial
generalized
mixed-
effects
model

Outcome: visit count
Main effect: group (RO vs. RWL)
Interaction term: group x instance
Other fixed effects: proficiency, memory, summed TRT
Random effects: person, item

$$\begin{aligned} \text{Visit Count} \sim & I + \\ & \text{Group} * \text{Instance} + \\ & \text{Proficiency} + \text{PSTM} + \\ & \text{SRT} + \text{summedTRT} + (I | \\ & \text{Person}) + (I | \text{Item}) + \\ & (\text{Instance} | \text{Person}) + \\ & (\text{summedTRT} | \text{Person}) + \\ & (\text{Group} | \text{Item}) + \\ & (\text{Instance} | \text{Item}) + \\ & (\text{Proficiency} | \text{Item}) + \\ & (\text{PSTM} | \text{Item}) + (\text{SRT} | \\ & \text{Item}) + (\text{summedTRT} | \\ & \text{Item}) + (\text{Group} : \text{Instance} \\ & | \text{Item}) \end{aligned}$$

1b

3-5

Generalized
multilevel
mixed
effects
models

Outcomes: form recognition, meaning recognition, form-
meaning recall
Main effect: group (RO vs. RWL)
Interaction term: group x item modality
Other fixed effects: proficiency, PSTM, SRT
Random effects: person, item

(1) *Form recognition
outcome*

$$\begin{aligned} \text{Form recognition} \sim & I + \\ & \text{Group} * \text{Item Modality} + \\ & \text{Proficiency} + \text{PSTM} + \\ & \text{SRT} + \text{Group} : \text{PSTM} + (I | \\ & \text{Person}) + (I | \text{Item}) + \\ & (\text{Item Modality} | \text{Person}) \\ & + (\text{Group} | \text{Item}) + \\ & (\text{Proficiency} | \text{Item}) + \\ & (\text{PSTM} | \text{Item}) + (\text{SRT} | \\ & \text{Item}) + (\text{Group} : \text{PSTM} | \\ & \text{Item}) \end{aligned}$$

(2) *Meaning recognition
outcome*

*Meaning recognition ~ I
+ Group*Item Modality +
Proficiency + PSTM +
SRT + Group:PSTM + (I
| Person) + (I | Item) +
(Item Modality | Person)
+ (Group | Item) +
(Proficiency | Item) +
(PSTM | Item) + (SRT |
Item) + (Group:PSTM |
Item)*

(3) *Meaning recall
outcome*

*Meaning recall ~ I +
Group*Item Modality +
Proficiency + PSTM +
SRT + Group:PSTM + (I
| Person) + (I | Item)
(Item Modality | Person)
+ (Group | Item) +
(Proficiency | Item) +
(PSTM | Item) + (SRT |
Item) + (Group:PSTM |
Item)*

2a	6	Linear multilevel mixed effects	Outcome: summed total reading time Main effect: proportion of reading ahead (within RWL) Interaction term: none Other fixed effects: proficiency, memory Random effects: person, item	$summedTRT \sim 1 +$ $Readahead + Proficiency$ $+ PSTM + SRT + (1 $ $Person) + (1 Item) +$ $(Readahead Person) +$ $(Proficiency Item) +$ $(PSTM Item) + (SRT $ $Item)$
2b	7-9	Linear multilevel mixed effects	Outcomes: form recognition, meaning recognition, form- meaning recall Main effect: proportion of reading ahead (within RWL) Interaction term: none Other fixed effects: proficiency, PSTM, SRT Random effects: person, item	(1) $Form\ recognition \sim 1$ $+ Readahead +$ $Proficiency + PSTM$ $+ SRT + (1 Person)$ $+ (1 Item)$ $(Readahead $ $Person) +$ $(Proficiency Item)$ $+ (PSTM Item) +$ $(SRT Item)$ (2) $Meaning\ recognition$ $\sim 1 + Readahead +$ $Proficiency + PSTM$ $+ SRT + (1 Person)$ $+ (1 Item)$ $(Readahead $ $Person) +$ $(Proficiency Item)$ $+ (PSTM Item) +$ $(SRT Item)$

(3) *Form-meaning recall*
 $\sim 1 + \text{Readahead} +$
 $\text{Proficiency} + \text{PSTM}$
 $+ \text{SRT} + (1 | \text{Person})$
 $+ (1 | \text{Item})$
 $(\text{Readahead} |$
 $\text{Person}) +$
 $(\text{Proficiency} | \text{Item})$
 $+ (\text{PSTM} | \text{Item}) +$
 $(\text{SRT} | \text{Item})$

3

10

See 1b
above

Outcomes: form recognition, meaning recognition, form-meaning recall
 Main effect: group (RO vs. RWL)
 Interaction term: group x PSTM
 Other fixed effects: proficiency, memory
 Random effects: person, item

See 1b above

Chapter 4: Results

4.1 Introduction to Results

In this chapter, results from the analyses are presented. Initial descriptive results from the individual difference variables (proficiency and PSTM) are given, along with the reading comprehension questions as evidence of a focus on meaning during the story reading task. Reliability estimates for outcomes and predictor variables are then provided, followed by a summary of the results from the data analyses to answer each research question. First, continuous eye-tracking results for gaze duration (GD), total reading time (TRT), and visit count are reported, through both descriptive and inferential statistics, to compare reading patterns by group (RO vs. RWL) and answer RQ1a. Secondly, dichotomous outcome analyses of vocabulary posttests are summarized through both descriptive and inferential statistics, to make group-level comparisons on the effects of RWL on learning outcomes, along with the impact of test item modality on learning (RQ1b). Additionally, these analyses included individual difference variables as predictors of outcomes to examine the unique impact of phonological short-term memory (PSTM) on learning new words under incidental conditions, and whether those would differ by group (RQ3). Third, results from both continuous and dichotomous mixed-effects models that operationalize and test reading ahead of the audio as a predictor of TRT and vocabulary learning outcomes within the RWL group (RQs 2a and 2b) are reported. Finally, qualitative data analysis of study debriefing responses is summarized.

4.2 Individual Difference Descriptives

Table 9 indicates mean score percentages for the two main proficiency measures (cloze and auditory LexTALE), along with raw descriptives for reading speed (in words per minute),

the two PSTM measures (running memory span and nonword span), and the measure of general processing speed (serial reaction task, or SRT). The two groups were remarkably similar across all tasks, with the RO group averaging slightly faster (24ms) in reading speed, while the RWL group averaged slightly faster in general processing (SRT) speed (8ms).

Descriptive Statistics for Individual Difference Variables in the Study

Table 9

Descriptive Statistics for Individual Difference Variables in the Study

Variable	RO Group (<i>n</i> = 60)					RWL Group (<i>n</i> = 59)				
	Mean	SD	Min	Max	95% CI	Mean	SD	Min	Max	95% CI
Cloze (%)	77.3	9.08	56	94	59.1, 95.2	79.5	4.37	60	96	70.76, 88.2
Auditory LexTALE (%)	67	10.6	42.5	83.75	45.8, 88.2	66.9	11.2	37.5	88.75	44.5, 89.3
Reading speed (WPM)	249	78.3	125	551	92.4, 405.6	225	62.8	125	458	99.4, 350.6
RMS (PSTM)	3.37	0.6	2.15	4.65	2.17, 4.57	3.38	0.7	2.05	5.05	1.98, 4.78
NWS (PSTM)	173.7	14.1	137	204	145.5, 201.9	174.5	17.5	106	206	139.5, 209.5
SRT (processing)	467.9	105.1	320.88	819.56	257.7, 678.1	459.9	99.8	288.56	770.63	260.3, 659.5

4.3 Story Comprehension Scores

As mentioned under Methods, comprehension questions were provided to participants prior to and following each chapter in order to ensure that the focus of the reading activity was on understanding the meaning of the story. 27 total questions (three per chapter) were used (see Appendix C). Overall scores for the comprehension questions reflected extremely high accuracy

($M = 0.964$; $SD = 0.187$; $Min = 0.778$; $Max = 1$), and nearly half of all participants (23 in the RO group; 32 in the RWL group) scored 100% on the comprehension questions. This was considered very strong evidence that participants were focused on story meaning during the reading task, so all participant datasets were kept.

4.4 Reliability Estimates

Table 10 indicates the linguistic and covariate measures employed, type of reliability reported, and reliability statistics. Overall reliability was considered acceptable across all measures, although the reliability of the Rasch IRT analysis of the meaning recall learning outcome indicated lower reliability (.51). Upon further examination, the low overall outcome scores across participants in the meaning recall measure (see descriptive scores in Table 21) indicated that some items did not differentiate well between participants using this method (i.e., no participant correctly answered them). As a result, the Kuder-Richardson 20 formula (Kuder & Richardson, 1937; Brennan, 2006) was employed as a further indicator of internal consistency for dichotomous data. The KR-20 formula for the meaning recall measure was .86, indicating stronger internal consistency. Further investigation of the effect of individual items on the measure may shed additional light on the low reliability estimates, but the measurement error is acknowledged.

Table 10

Reliability Estimates for Each Vocabulary Learning Outcome and Covariate Measure

Measure	Type of reliability	Coefficient(s)
Form recognition	Rasch IRT person reliability / KR-20	Rasch IRT: .69 KR-20: .76
Meaning recognition	Rasch IRT person reliability / KR-20	Rasch IRT: .64 KR-20: .66

Meaning recall	Rasch IRT person reliability / KR-20	Rasch IRT: .51 KR-20: .86
Cloze reading ability	Rasch IRT person reliability / KR-20	Rasch IRT: .65 KR-20: .67
Auditory LexTALE	Rasch IRT person reliability / KR-20	Rasch IRT: .72 KR-20: .74
Running memory span	Cronbach's α	.75
Nonword span	Cronbach's α	.90
Serial reaction time (processing)	Split half (Spearman-Brown correction)	.98

4.5 Eye-Tracking Measures

4.5.1 Graphs and descriptive statistics

Prior to inferential analyses, graphs of eye-tracking data for GD, TRT, and visit count across the ten encounters with each pseudoword were created for visual inspection. These consisted of plots of means, fitted loess lines (smoothers), and 95% confidence intervals (shaded in gray around the lines). I did this to visualize general patterns of change across time as participants encountered each new word's 10 instances in the text, and to examine the raw data for the predicted S-shaped longitudinal patterns in GD and TRT. Graphs of raw means for GD, TRT, and visit count across all instances of target items are included by group in Figures 3-5, while exact means, standard deviations, and 95% CIs are provided in Tables 11-13. Observations where targets were skipped were excluded from the dataset prior to calculating these means and standard deviations. This resulted in the removal of 855 of the 29,631 total data points, or 2.9% of the total eye-tracking data.

For GD, an early measure of lexical access (Godfroid, 2020a), there was an initial difference of 38ms between the two groups, with the RO group spending less initial time at the initial encounter with each new word than in the RWL group. Interestingly, while the RO group demonstrated a noticeable decrease in GD across instances of the new words in the text, there was a noticeably smaller change in GD across instances of the new words in the RWL group, and the RO group consistently spent less initial time reading the new words in the text.

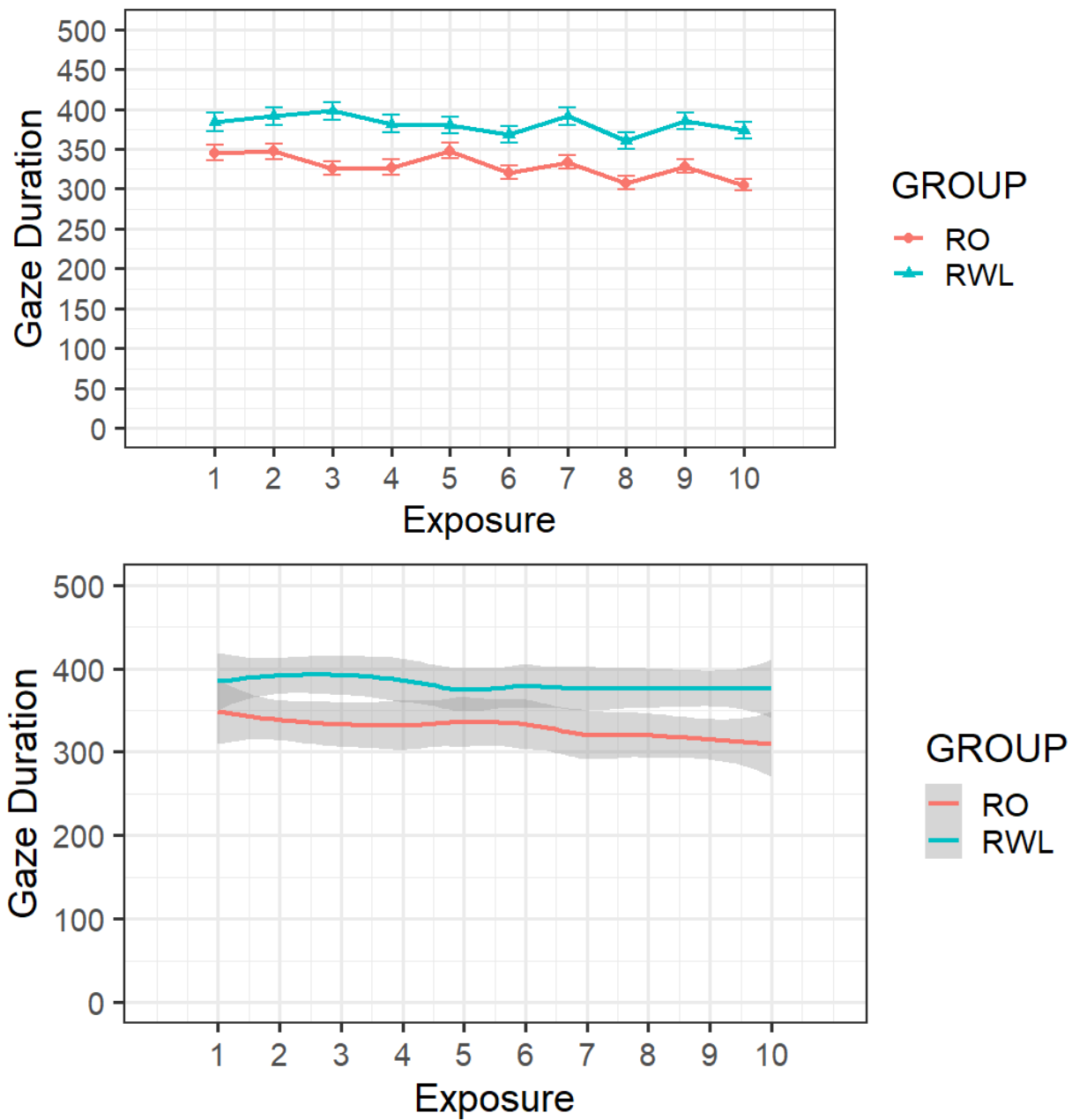


Figure 3. Data visualization for gaze duration ET measure by group. (first plot is means; second plot is fitted loess lines with 95% confidence intervals).

Table 11*Means, Standard Deviations, and 95% Confidence Intervals for Gaze Duration by Group*

Instance	Reading Only				Reading While Listening			
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>95%CIs</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>95%CIs</i>
1	1442	345.19	193.63	[335, 355]	1446	383.93	220.85	[372, 394]
2	1470	347.21	187.00	[337, 357]	1454	391.38	213.81	[380, 402]
3	1536	325.56	171.12	[317, 335]	1510	397.62	226.12	[387, 409]
4	1441	326.97	182.45	[318, 336]	1442	381.46	215.41	[370, 392]
5	1450	348.02	188.31	[338, 358]	1428	379.78	206.79	[369, 391]
6	1443	320.59	169.52	[312, 330]	1439	368.52	198.94	[359, 379]
7	1465	333.38	167.11	[324, 342]	1386	391.13	216.34	[380, 402]
8	1429	307.72	158.61	[300, 316]	1376	360.33	190.35	[350, 370]
9	1392	328.23	163.67	[319, 337]	1431	385.49	205.94	[374, 396]
10	1368	304.98	139.59	[298, 312]	1428	373.62	205.23	[363, 385]

Visit count data initially exhibited the opposite pattern from gaze duration, and are represented in Figure 4 and Table 12. Participants in the RO group visited the target interest areas more frequently and with greater variability across instances in the text than in RWL, with a general decreasing trend in visit count for both groups. In the instance 6-7 range, the scores for the two groups began overlapping to a greater degree (see Table 12 for CIs by instance), indicating that this trend was sharper in RO than RWL, but variability in visit count remained noticeably higher in in RO across instances.

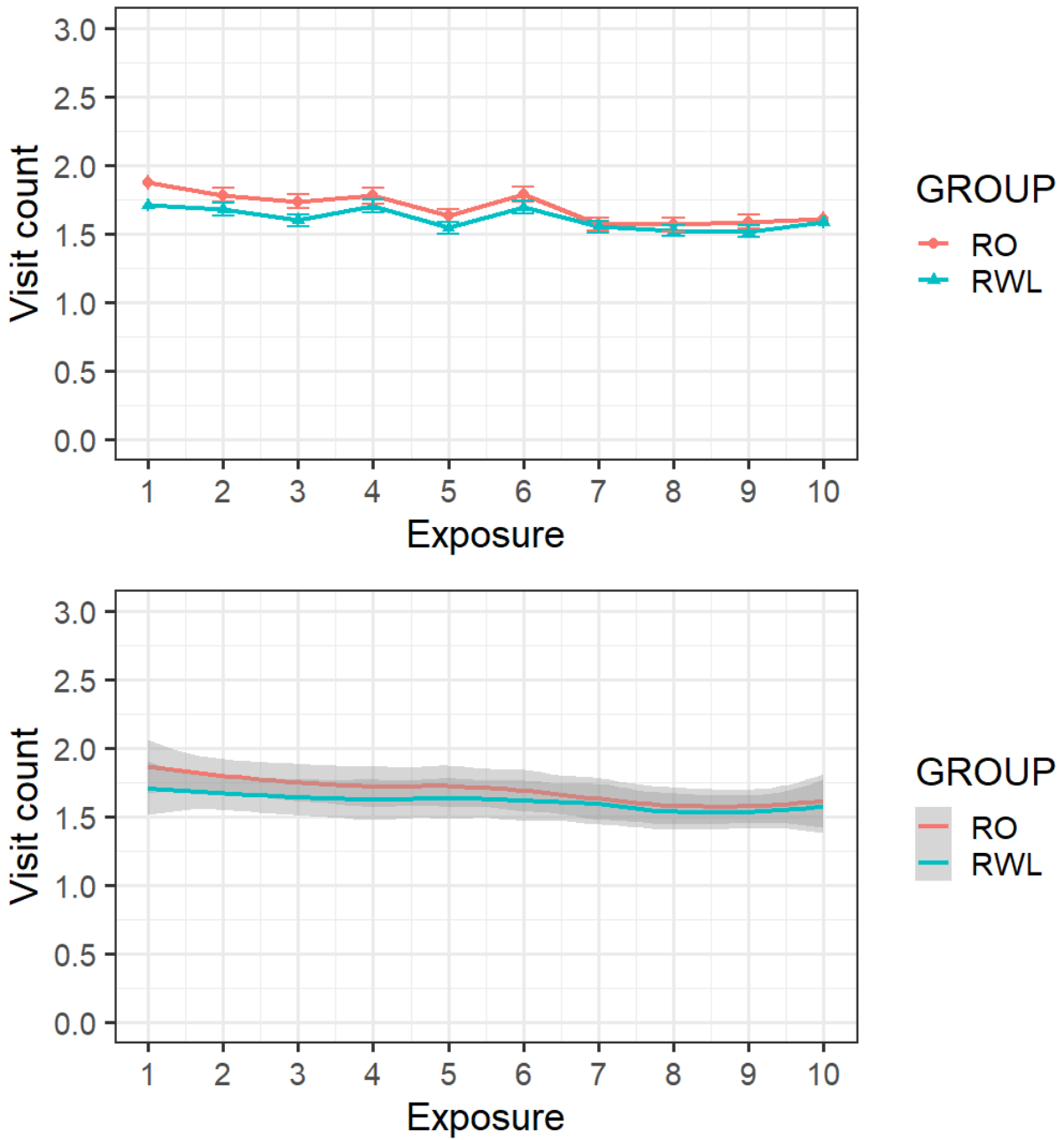


Figure 4. Data visualization for visit count ET measure by group. (first plot is means; second plot is fitted loess lines with 95% confidence intervals).

Table 12*Means, Standard Deviations, and 95% Confidence Intervals for Visit Count by Group*

Instance	Reading Only				Reading While Listening			
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>95%CIs</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>95%CIs</i>
1	1442	1.87	1.07	[1.81, 1.93]	1446	1.71	0.88	[1.66, 1.76]
2	1470	1.78	1.01	[1.73, 1.83]	1454	1.68	0.88	[1.63, 1.73]
3	1536	1.74	1.03	[1.69, 1.79]	1510	1.60	0.85	[1.56, 1.64]
4	1441	1.78	1.10	[1.72, 1.84]	1442	1.70	0.93	[1.65, 1.75]
5	1450	1.63	0.93	[1.58, 1.68]	1428	1.54	0.76	[1.50, 1.58]
6	1443	1.79	1.10	[1.73, 1.85]	1439	1.70	0.94	[1.65, 1.75]
7	1465	1.57	0.92	[1.52, 1.62]	1386	1.55	0.81	[1.51, 1.59]
8	1429	1.57	0.94	[1.52, 1.62]	1376	1.52	0.77	[1.48, 1.56]
9	1392	1.59	0.93	[1.54, 1.64]	1431	1.52	0.80	[1.48, 1.56]
10	1368	1.61	1.03	[1.56, 1.66]	1428	1.59	0.94	[1.54, 1.64]

The GD and visit count data informed understanding of the visuals and descriptive statistics for TRT, demonstrated by Figure 5 and Table 13, since both GD and visit count are eye tracking measures interdependent with TRT (e.g., Godfroid, 2020b). The two groups were nearly identical in TRT during initial instances, reflecting balance between higher GD times for RWL participants and more visits to the interest areas with target items for RO participants, during initial encounters. Both groups exhibited a gradual decrease in TRT, but the RO group times decreased more sharply, and with greater variability, than the more steady and gradual decrease in TRT in RWL. The difference between the groups more clearly emerged between instances 6 and 7, with mean times clearly faster in RO (on average, 30-50 ms faster). However, as Table 13 indicates, variability was also much higher across all instances for participant TRT in RO.

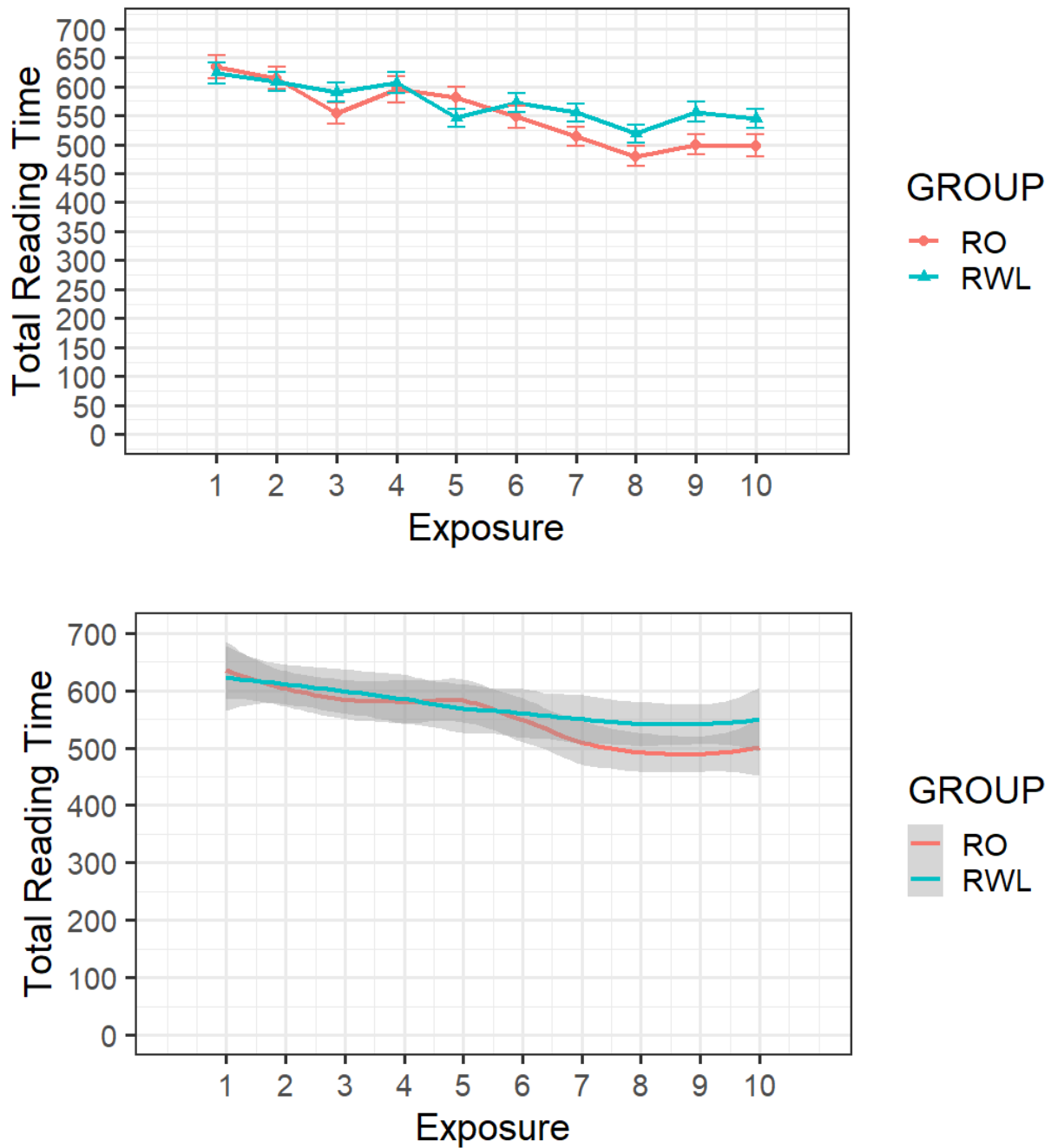


Figure 5. Data visualization for total reading time ET measure by group. (first plot is means; second plot is fitted loess lines with 95% confidence intervals).

Table 13*Means, Standard Deviations, and 95% Confidence Intervals for Total Reading Time by Group*

Instance	Reading Only				Reading While Listening			
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>95%CIs</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>95%CIs</i>
1	1442	633.78	395.79	[613, 653]	1446	623.65	353.26	[606, 642]
2	1470	614.59	379.62	[596, 634]	1454	607.86	320.54	[592, 624]
3	1536	553.74	371.26	[534, 582]	1510	590.12	334.23	[573, 607]
4	1441	595.11	445.34	[572, 618]	1442	606.35	356.00	[588, 624]
5	1450	580.36	374.81	[561, 599]	1428	545.60	292.58	[531, 560]
6	1443	547.54	384.96	[528, 568]	1439	571.98	331.84	[555, 589]
7	1465	513.31	326.28	[496, 530]	1386	555.15	292.18	[540, 570]
8	1429	479.91	329.26	[463, 497]	1376	518.28	300.38	[502, 534]
9	1392	500.00	342.01	[482, 518]	1431	556.13	330.19	[539, 573]
10	1368	498.18	371.33	[478, 518]	1428	545.04	311.48	[529, 561]

Following basic descriptives, I examined the correlations between the three eye-tracking measures (GD, TRT, visit count), given assumptions that they are strongly related to one another (e.g., Godfroid, 2020b). I standardized mean values for each metric by participant using z-transformations to improve normality assumptions and make meaningful comparisons, and ran Pearson correlation coefficients to determine the size of the relationships. Table 14 summarizes them, with moderately strong relationships and R squared values reported. As expected, there was a positive relationship between GD and TRT ($r = 0.56$, $R^2 = 0.32$), and between visit count and TRT ($r = 0.35$, $R^2 = 0.12$), and a negative relationship between GD and visit count ($r = -0.53$, $R^2 = 0.28$). In other words, as GD and visit count increased, so did TRT across participants, while as GD increased, visit count decreased.

Table 14*Correlation Coefficients for Relationships between Eye-Tracking Variables*

	Gaze Duration	Visit Count	Total Reading Time
Gaze Duration	N/A	Pearson's $r = -0.53$ $R^2 = 0.28$	Pearson's $r = 0.56$ $R^2 = 0.32$
Visit Count	Pearson's $r = -0.53$ $R^2 = 0.28$	N/A	Pearson's $r = 0.35$ $R^2 = 0.12$
Total Reading Time	Pearson's $r = 0.56$ $R^2 = 0.32$	Pearson's $r = 0.35$ $R^2 = 0.12$	N/A

4.5.2 Inferential statistics for between-groups eye-tracking data comparisons

Following the examination of descriptive statistics and visualization of the data, linear mixed-effects models were fitted to examine statistical significance in the difference between groups (RO vs. RWL) on the three eye-tracking variables of interest – GD, TRT, and visit count. These were fit to answer RQ1a regarding differences in reading patterns between RO and RWL, and in the initial dissertation proposal it was predicted that there would be (1) a significant decrease in GD and TRT for both RO and RWL; (2) longer GD and TRT and higher visit count in RO than RWL. Models were fitted separately for each measure, with a backward stepwise approach to model fitting, initially including all covariates. Model fitting proceeded with a maximal model, with non-significant effects of covariate predictors removed until arriving at the best-fitting model. I utilized chi-square difference tests for the growth curve models using the `anova()` function in R. The models utilizing quadratic and cubic terms did not add to model fit, so the final reported models were those for GD and TRT that included only the linear term. Model comparisons for simple, maximal (not including random slopes), and best-fitting models are provided in Appendix L, while the best-fitting models are now summarized.

The best-fitting model for GD is provided in Table 15 and visualized in Figure 6. Data visualization and recent similar studies examining L2 eye-tracking while reading (e.g., Godfroid et al., 2018; Tuzcu, 2023) found a non-linear decrease in GD across initial exposures to new words, so I expected to find the same pattern. However, the best-fitting model for the GD data in the present study did not include quadratic or cubic terms, but was *linear* in nature. There were significant main effects of group ($b = 0.09$, $SE = 0.03$, $p = .002$), instance both for RO ($b = -0.01$, $SE = 0.00$, $p < .001$) and RWL ($b = -0.004$, $SE = 0.001$, $p = .004$), and a group by instance interaction ($b = 0.00$, $SE = 0.00$, $p = .011$). Additionally, proficiency ($b = -0.13$, $SE = 0.02$, $p < .001$) was a significant negative covariate predictor, while PSTM ($b = 0.04$, $SE = 0.02$, $p = .038$) and visit count ($b = -0.07$, $SE = 0.00$, $p < .001$) were significant positive covariate predictors, and were kept in the best-fitting model for GD. To summarize main findings, Participants in the RO group exhibited *shorter* GD times across instances compared with the RWL group (different than predicted), both groups significantly decreased in GD across instances (as predicted), and the decrease across instances was greater in the RO group than the RWL group (different than predicted).

Table 15

Best-fitting Model for Gaze Duration

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>t</i>	<i>p</i>	<i>Exp(b)</i>
Intercept	5.85	0.03	5.80 – 5.90	229.182	<.001	346.19
Group (RWL) ¹	0.09	0.03	0.03 – 0.14	3.128	.002	1.09
Instance (Linear)	-0.01	0.00	-0.01 – -0.01	-6.588	<.001	0.99
Group (RWL) x Instance (Linear)	0.00	0.00	0.00 – 0.01	2.53	.011	1.00
<i>Covariates</i>						
Proficiency	-0.13	0.02	-0.17 – -0.10	-6.944	<.001	0.88
PSTM	0.04	0.02	0.00 – 0.07	2.075	.038	1.04

Visit Count	-0.07	0.00	-0.08 - -0.07	-23.802	<.001	0.93
<i>Random Effects</i>		<i>Variance</i>	<i>SD</i>			
1 Participant	0.02	0.14				
1 Item	0.01	0.07				
Residual	0.21	0.46				
N _{participant}	119					
N _{item}	25					
Observations	28246					
Marginal R ² /	0.073 /					
Conditional R ²	0.174					

¹The reading-only group is the baseline for this summary.

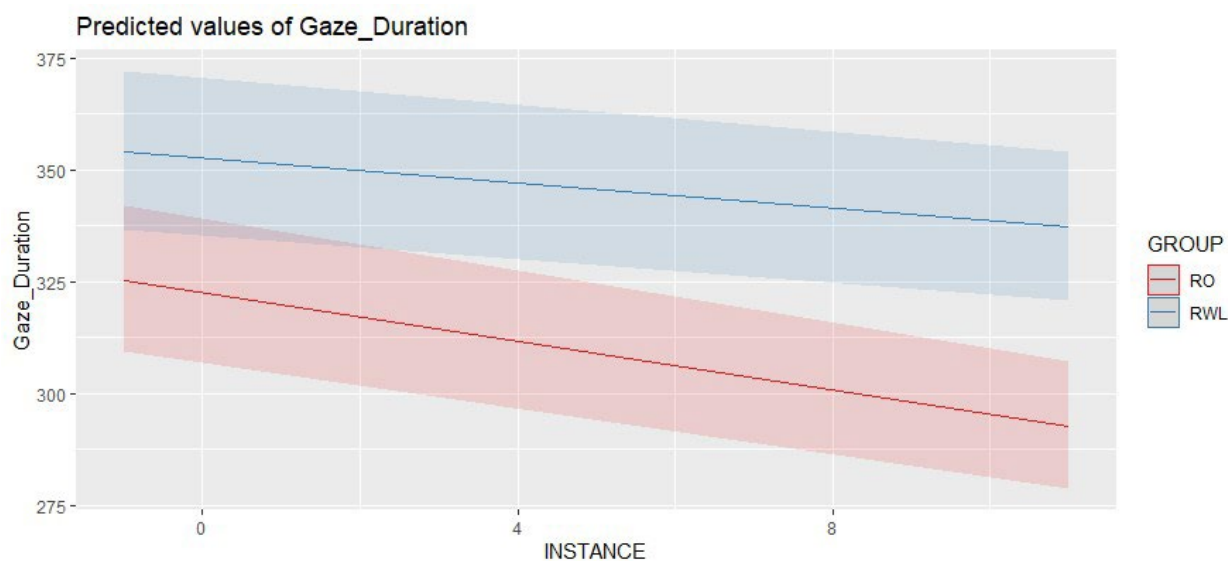


Figure 6. Best-fitting model-based estimates of gaze duration by group across instances.

The best-fitting model for TRT is provided in Table 16 and visualized in Figure 7. As with the GD data, the best-fitting model for the TRT data did not include quadratic or cubic terms, but was *linear* in nature. There were significant main effects of group ($b = 0.05$, $SE = 0.03$, $p = .046$) and instance both for RO ($b = -0.02$, $SE = 0.001$, $p < .001$) and RWL ($b = -0.01$, $SE = 0.001$, $p < .001$), and a significant interaction between group and instance ($b = 0.10$, $SE = 0.002$, $p < .001$). Proficiency ($b = -0.08$, $SE = 0.02$, $p < .001$) was a significant negative covariate

predictor in the model, while PSTM ($b = 0.04$, $SE = 0.02$, $p = .02$) and visit count ($b = 0.27$, $SE = 0.03$, $p < .001$) were significant positive covariate predictors of TRT. To summarize findings briefly, participants in the RO group exhibited *shorter* TRT across instances than the RWL group (different than predicted), both groups significantly decreased in TRT across instances (as predicted), and the decrease across instances was greater in the RO than the RWL group (different than predicted).

Table 16

Best-fitting Model for Total Reading Time

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>t</i>	<i>p</i>	<i>Exp(b)</i>
Intercept	5.79	0.02	5.74 – 5.84	236.167	<.001	326.69
Group (RWL) ¹	0.05	0.03	0.00 – 0.10	1.998	.046	1.05
Instance (Linear)	-0.02	0.001	-0.02 – -0.02	-15.633	<.001	0.98
Group (RWL) x Instance (Linear)	0.10	0.002	0.096 – 0.104	4.951	<.001	1.11
<i>Covariates</i>						
Proficiency	-0.08	0.02	-0.11 – -0.04	-4.599	<.001	0.92
PSTM	0.04	0.02	0.00-0.08	2.336	.02	1.04
Visit Count	0.27	0.03	0.21-0.33	84.516	<.001	1.31
<i>Random Effects</i>						
1 Participant	0.02	0.12				
1 Item	0.01	0.08				
Residual	0.22	0.47				
N _{participant}	119					
N _{item}	25					
Observations	28246					
Marginal R ² / Conditional R ²	0.221 / 0.288					

¹The reading-only group is the baseline for this summary.

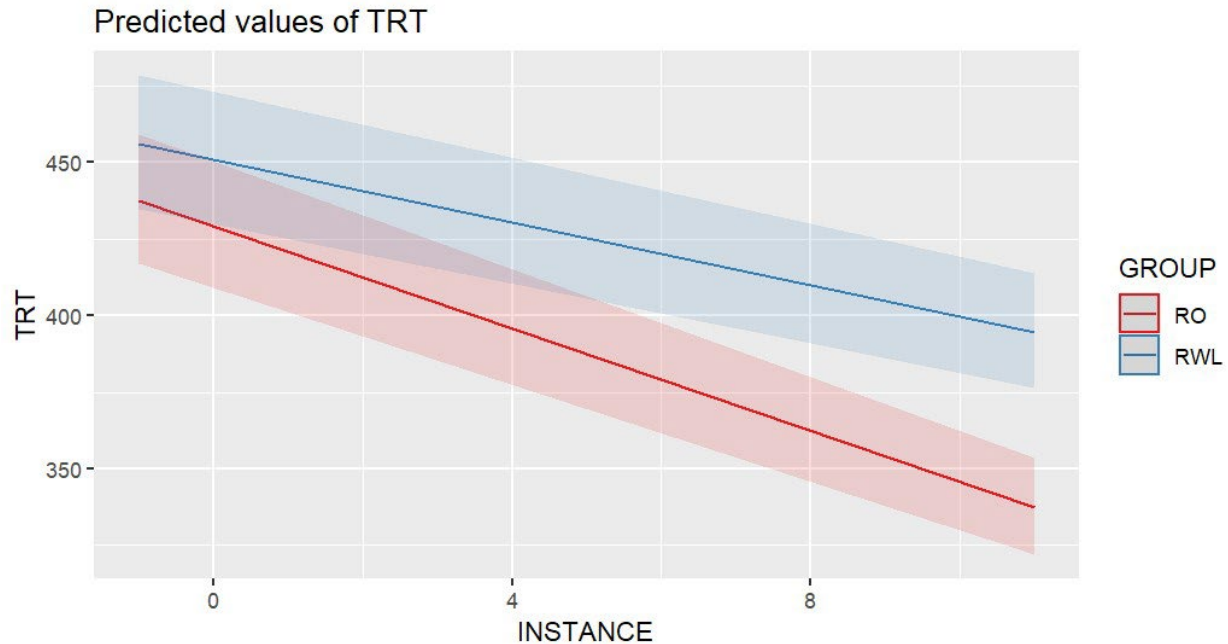


Figure 7. Best-fitting model-based estimates of TRT across instances.

For the analysis of visit count, a negative binomial logistic mixed-effects regression model was built, with the best-fitting model summary for visit count provided in Table 17 and visualized in Figure 8. Initial models were tested in a backward stepwise fashion, with non-significant predictors removed to improve model fit in the same way as for the GD and TRT analyses. Initial maximal models did not converge for the visit count data, so I mostly resolved nonconvergence and singular fit issues with the visit count data by utilizing the *bobyqa* optimizer from the *lme4* package in R (R Core Team, 2023), to increase the number of model iterations to arrive at convergence. As a result, the best-fitting model shown in Table 17 converged. Models that included SRT and PSTM (together or separately) did not converge, or gave singularity warnings, so these variables were not included in the final models.

Significant main effects of group ($b = -0.08$, $SE = 0.03$, $p = .005$) and instance ($b = -0.02$, $SE = 0.002$, $p < .001$) were found for the visit count data, indicating that participants in the RO group made more visits to the target interest areas than those in the RWL group (as predicted),

and there was a significant decrease in visit count across instances for participants in both groups (as predicted). Additionally, the significant group by instance interaction ($b = 0.01$, $SE = 0.003$, $p = .031$) indicates that the decrease in visit count was sharper in the RO group than the RWL group (as predicted), which matches well with the data visualization of raw counts shown earlier in Figure 4. Proficiency was a significant positive predictor of visit count, as well ($b = 0.08$, $SE = 0.02$, $p < .001$).

Table 17

Best-fitting Negative Binomial Logistic Mixed-Effects Model for Visit Count

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>z</i>	<i>p</i>	<i>Exp(b)</i>
Intercept	0.61	0.03	0.55-0.67	23.087	<.001	1.84
Group (RWL) ¹	-0.08	0.03	-0.14 – -0.02	-2.833	.005	0.92
Instance	-0.02	0.002	-0.02 – -0.02	-7.771	<.001	0.98
<i>Covariates</i>						
Proficiency	0.08	0.02	0.04 – 0.12	4.874	<.001	1.08
Group (RWL) x Instance	0.01	0.003	0.01-0.01	2.152	.031	1.01
<i>Random Effects</i>						
1 Participant	0.01	0.12				
1 Item	0.01	0.08				
Residual	0.48	0.69				
N _{participant}	119					
N _{item}	25					
Observations	28246					
Marginal R ² / Conditional R ²	0.011 / 0.052					

¹The reading-only group is the baseline for this summary.

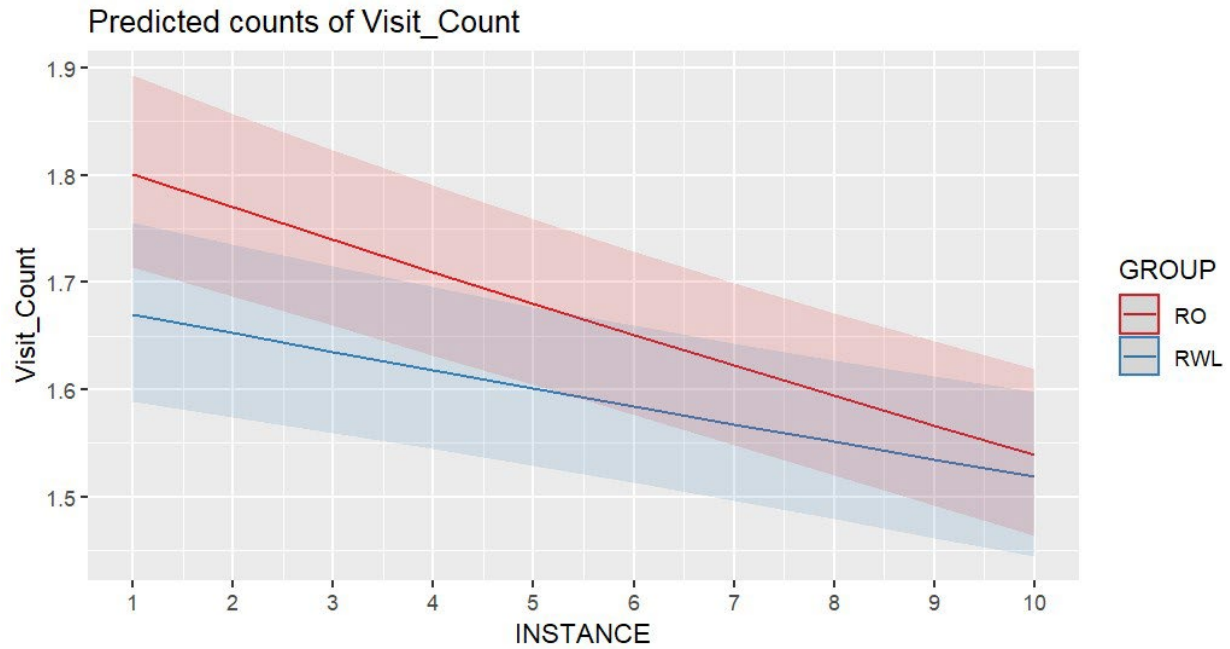


Figure 8. Best-fitting model-based estimates of visit count across instances.

4.5.3 Summary of results for between-groups eye-tracking comparisons

Table 18 summarizes the main findings of the between-groups analyses of the eye-tracking data, including relevant research questions, predictions, findings, and alignment of predictions with findings for each of the three measures (GD, TRT, and Visit Count).

Table 18

Summary of Predictions and Results from Eye-tracking Measures

<i>Research question</i>	<i>Eye-tracking Measure</i>	<i>Predictions</i>	<i>Findings</i>	<i>Alignment</i>
(1a) Is the online processing trajectory of form acquisition (as measured by eye tracking) for novel L2 words different under incidental conditions during reading while listening, compared with reading alone?	Gaze Duration	<ul style="list-style-type: none"> • A statistically-significant <i>curvilinear</i> decrease in GD for both groups 	<ul style="list-style-type: none"> • A statistically-significant <i>linear</i> decrease in GD for both groups 	Partial
		<ul style="list-style-type: none"> • A significant group by instance interaction, indicating a sharper decrease in GD for participants in the <i>RWL</i> group 	<ul style="list-style-type: none"> • A significant group by instance interaction, indicating a sharper decrease in GD for participants in the <i>RO</i> group 	X
		<ul style="list-style-type: none"> • Greater variation in GD for participants in the <i>RWL</i> than <i>RO</i> group 	<ul style="list-style-type: none"> • Greater variation in GD for participants in the <i>RWL</i> than <i>RO</i> group 	✓
		<ul style="list-style-type: none"> • Proficiency as a significant <i>negative</i> covariate predictor of GD 	<ul style="list-style-type: none"> • Proficiency as a significant <i>positive</i> covariate predictor of GD 	X

(1a) Is the online processing trajectory of form acquisition (as measured by eye tracking) for novel L2 words different under incidental conditions during reading while listening, compared with reading alone?	Total reading time	<ul style="list-style-type: none"> • A statistically-significant <i>curvilinear</i> decrease in TRT for both groups 	<ul style="list-style-type: none"> • A statistically-significant <i>linear</i> decrease in TRT for both groups 	Partial
		<ul style="list-style-type: none"> • A significant group by instance interaction, indicating a sharper decrease in TRT for participants in the <i>RWL</i> group 	<ul style="list-style-type: none"> • A significant group by instance interaction, indicating a sharper decrease in TRT for participants in the <i>RO</i> group 	X
		<ul style="list-style-type: none"> • Greater variation in TRT for participants in the <i>RO</i> group 	<ul style="list-style-type: none"> • Greater variation in TRT for participants in the <i>RO</i> group 	✓
	Visit count	<ul style="list-style-type: none"> • A statistically-significant decrease in visit count for both groups across instances 	<ul style="list-style-type: none"> • A statistically-significant decrease in visit count for both groups across instances 	✓
		<ul style="list-style-type: none"> • A significant difference between <i>RO</i> and <i>RWL</i> groups, indicating more visits for the <i>RO</i> group during early encounters with new words 	<ul style="list-style-type: none"> • A significant difference between <i>RO</i> and <i>RWL</i> groups, indicating more visits for the <i>RO</i> group during early encounters with new words 	✓
		<ul style="list-style-type: none"> • Greater variation in visit count for participants in the <i>RO</i> group 	<ul style="list-style-type: none"> • Greater variation in visit count for participants in the <i>RO</i> group 	✓

4.6 Vocabulary Learning Outcomes

4.6.1 Graphs and descriptive statistics

Tables 19-21 reveal raw mean scores by group and modality on the three vocabulary learning outcomes – form recognition, meaning recognition, and meaning recall. Since participants were presented with each item randomly either in the visual or auditory modality, the k size for total items was similar between item modality and group, but not exactly equal. As such, Figures 9-17 demonstrate group and item modality comparisons for mean scores, rescaled to proportion of accurate responses by participant, group, and item modality for more interpretable visual comparison.

Table 19

Descriptive Statistics for Form Recognition Raw Learning Gains

	Reading Only			Reading While Listening		
	<i>n</i> (average)	<i>M</i> (<i>SD</i>)	95% <i>CIs</i>	<i>n</i> (average)	<i>M</i> (<i>SD</i>)	95% <i>CIs</i>
Visual items	12.68	9.81 (5.30)	[9.44, 10.19]	12.69	9.37 (5.58)	[8.97, 9.77]
Auditory items	12.32	8.32 (5.77)	[7.90, 8.73]	12.31	9.34 (5.27)	[8.96, 9.73]
All items	25	18.13 (5.53)	[17.34, 18.92]	25	18.71 (5.42)	[17.93, 18.50]

Note: Mean scores for target items only on form recognition test.

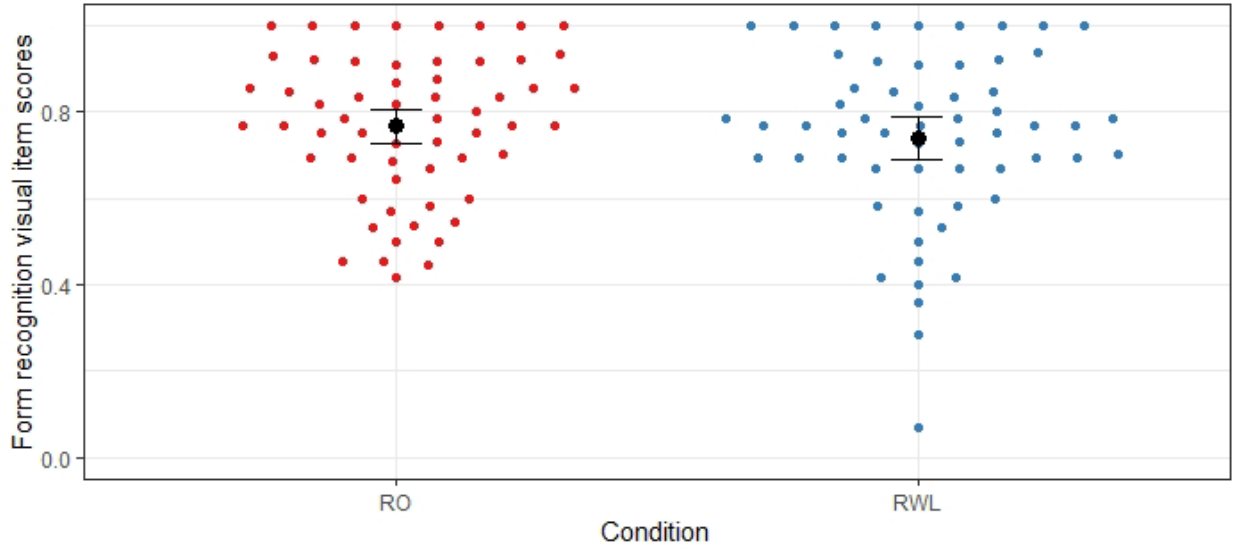


Figure 9. Bee swarm plot of mean score percentages and 95% confidence intervals by group on form recognition visual items.

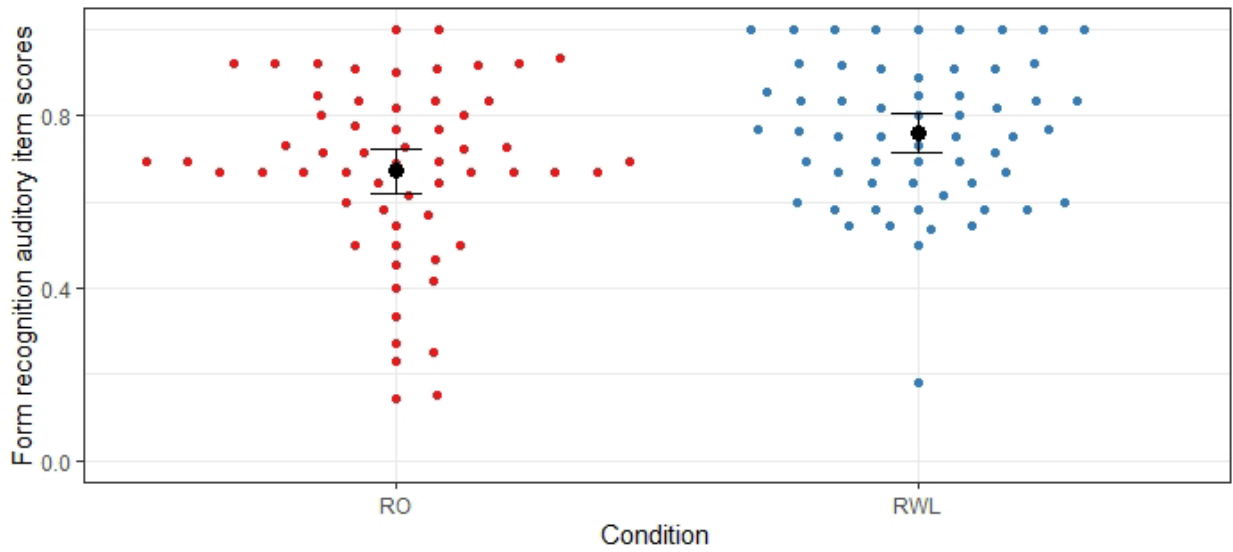


Figure 10. Bee swarm plot of mean score percentages and 95% confidence intervals by group on form recognition auditory items.

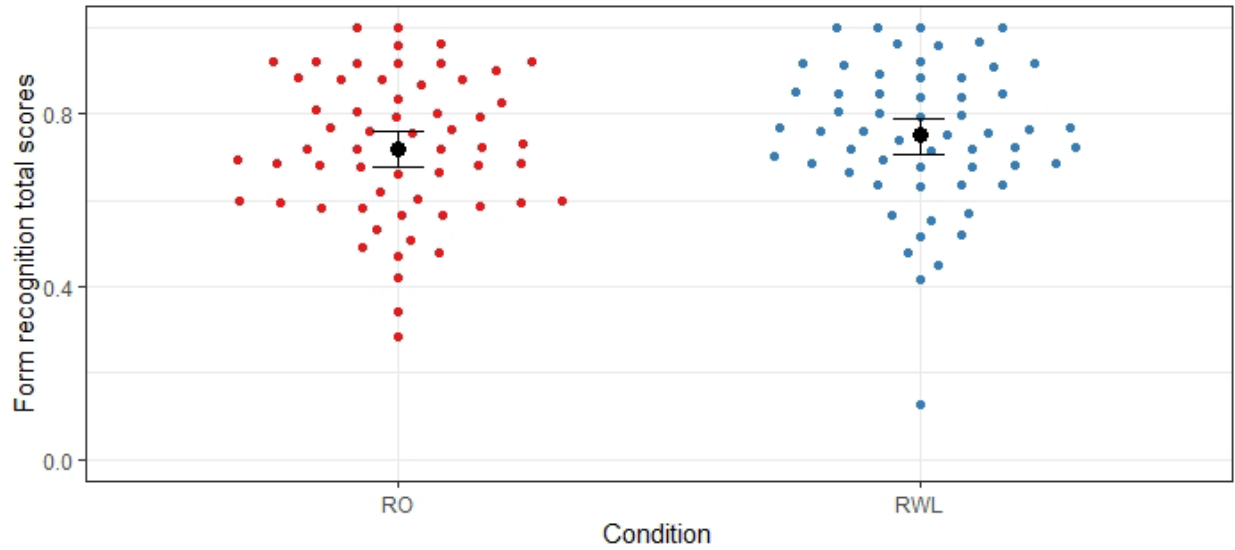


Figure 11. Bee swarm plot of mean score percentages and 95% confidence intervals by group on all form recognition items.

Table 20

Descriptive Statistics for Meaning Recognition Raw Learning Gains

	Reading Only			Reading While Listening		
	<i>n</i> (average)	<i>M</i> (<i>SD</i>)	95% <i>CIs</i>	<i>n</i> (average)	<i>M</i> (<i>SD</i>)	95% <i>CIs</i>
Visual items	12.38	7.03 (6.14)	[6.59, 7.47]	12.39	7.17 (6.12)	[6.73, 7.61]
Auditory items	12.62	6.50 (6.26)	[6.05, 6.95]	12.61	7.05 (6.26)	[6.60, 7.50]
All items	25	13.53 (12.46)	[12.90, 14.16]	25	14.22 (12.38)	[13.59, 14.85]

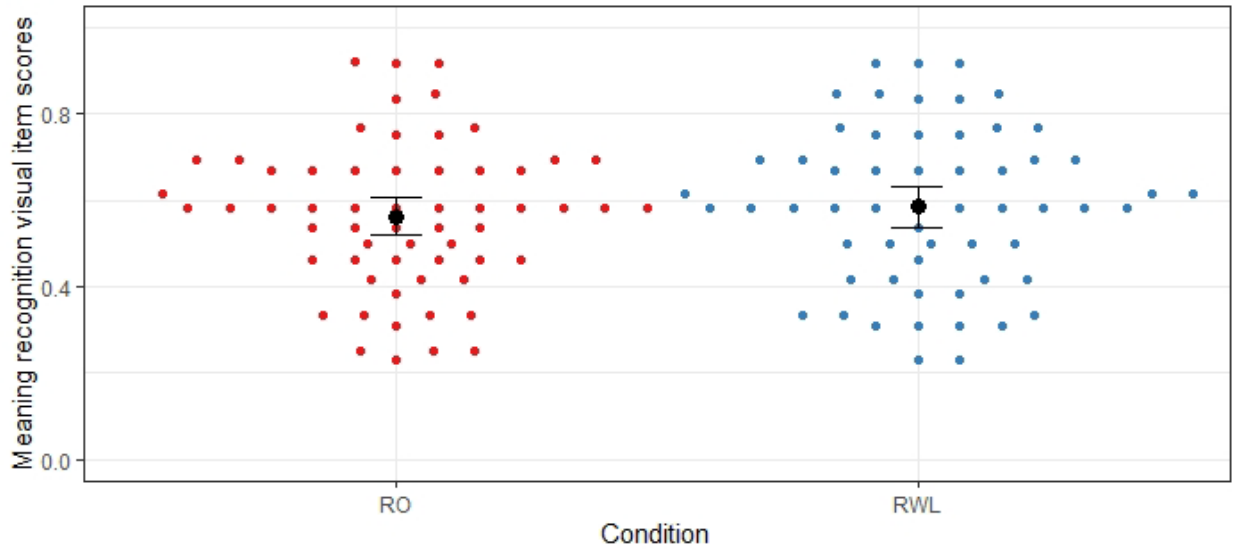


Figure 12. Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recognition visual items

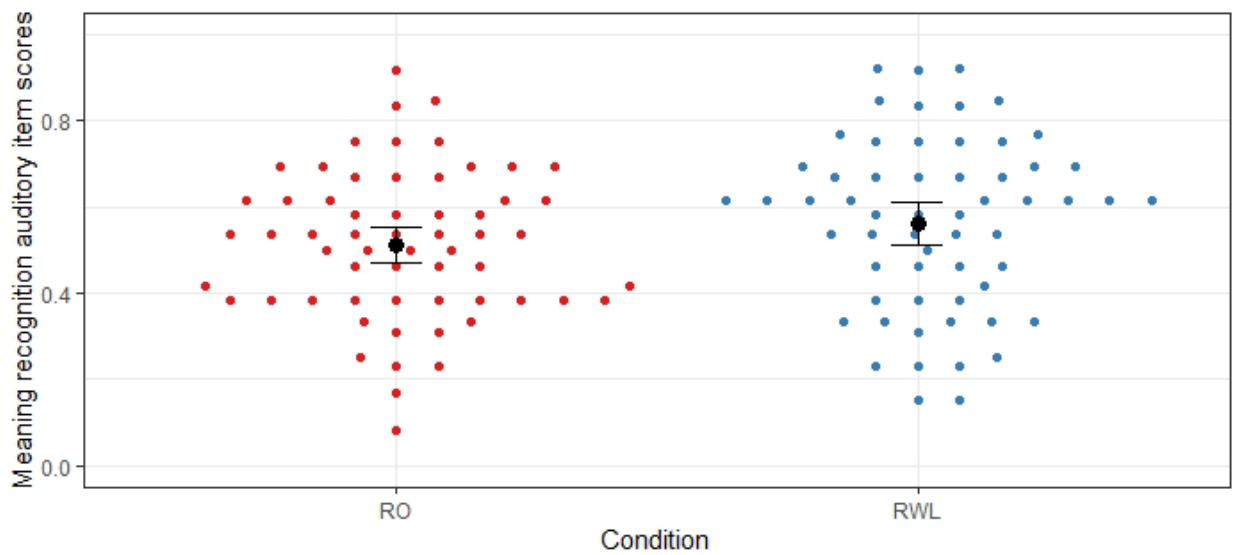


Figure 13. Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recognition auditory items.

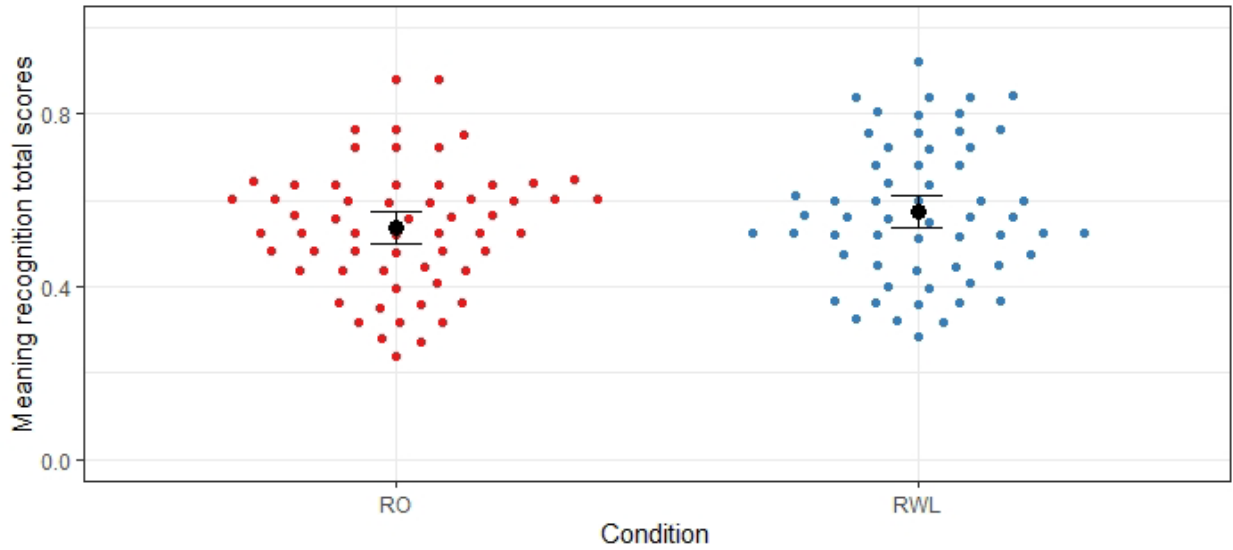


Figure 14. Bee swarm plot of mean score percentages and 95% confidence intervals by group on all meaning recognition items.

Table 21

Descriptive Statistics for Meaning Recall Raw Learning Gains

	Reading Only			Reading While Listening		
	<i>n</i> (average)	<i>M</i> (<i>SD</i>)	95% <i>CIs</i>	<i>n</i> (average)	<i>M</i> (<i>SD</i>)	95% <i>CIs</i>
Visual items	12.47	1.38 (3.92)	[1.10, 1.66]	12.47	2.02 (4.59)	[1.69, 2.35]
Auditory items	12.53	0.90 (3.24)	[0.67, 1.13]	12.53	2.10 (4.68)	[1.76, 2.44]
All items	25	2.28 (7.20)	[1.92, 2.64]	25	4.10 (9.28)	[3.63, 4.57]

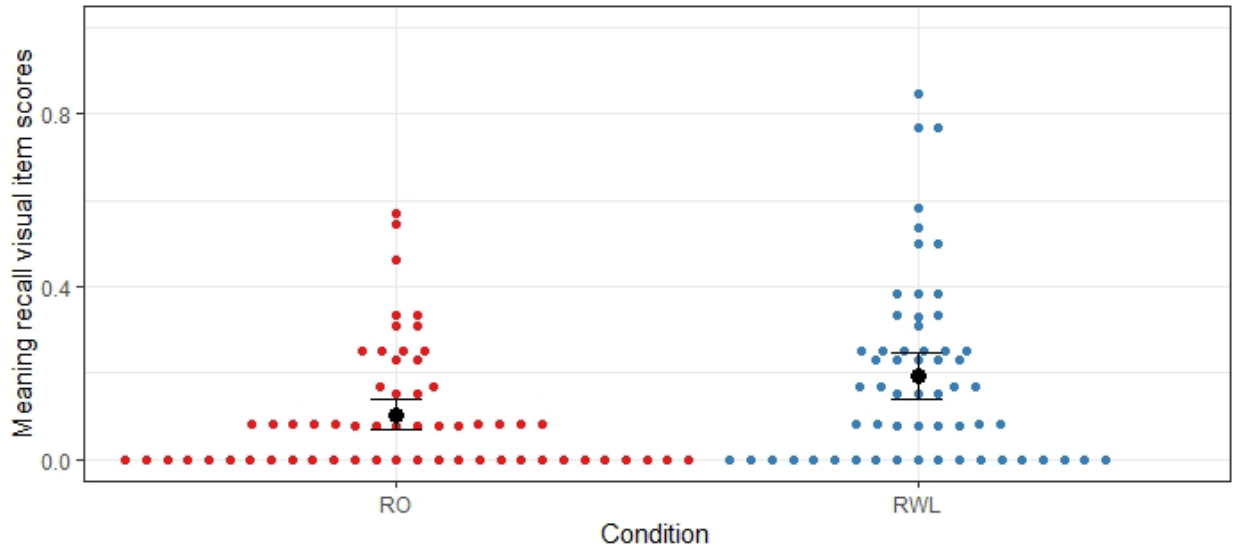


Figure 15. Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recall visual items.

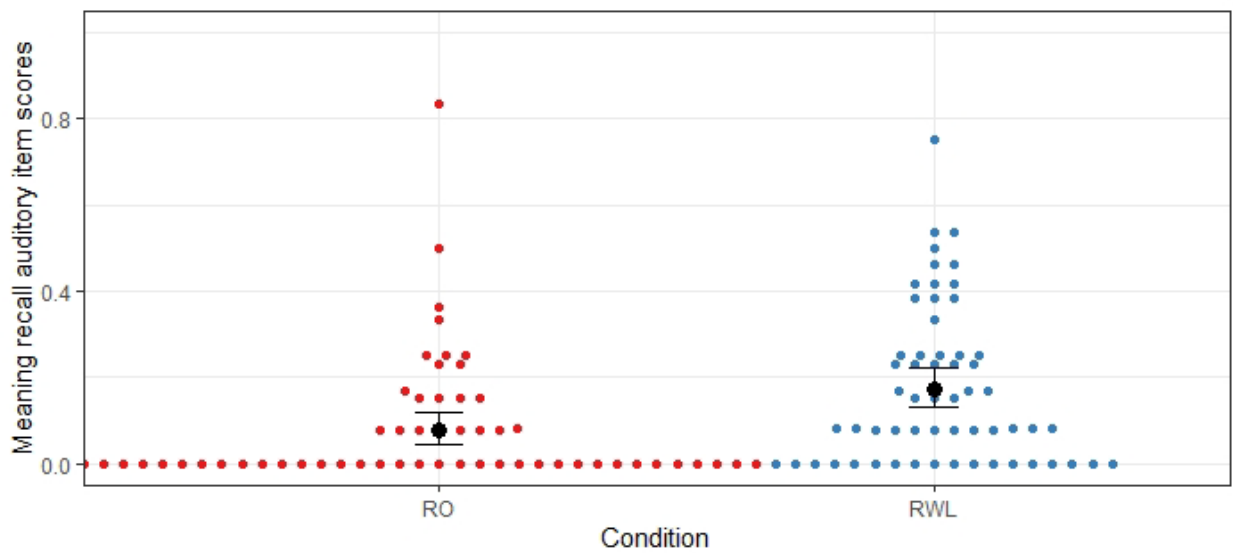


Figure 16. Bee swarm plot of mean score percentages and 95% confidence intervals by group on meaning recall auditory items.

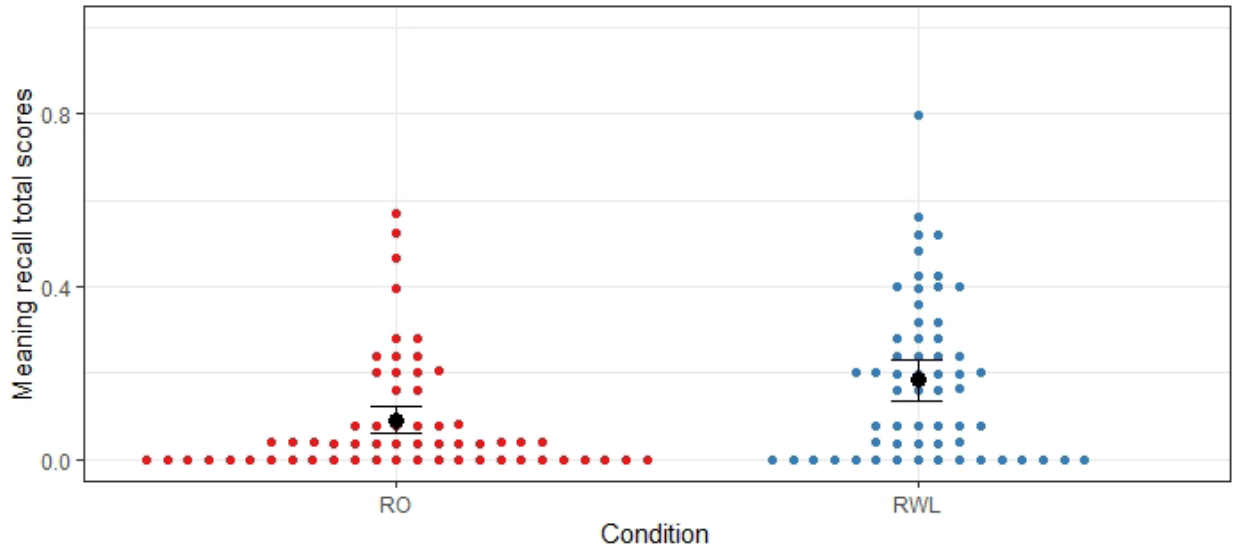


Figure 17. Bee swarm plot of mean score percentages and 95% confidence intervals by group on all meaning recall items.

4.6.2 Inferential statistics for between-group VL outcome comparisons

Logistic mixed-effects regression models were built for each of the three learning outcomes, with the main effects of group (RO vs. RWL) and item modality (visual vs. auditory) and their interaction as the primary categorical predictors for each outcome model. Composite proficiency and PSTM, along with SRT (general processing) and summed TRT, were included in the omnibus models as covariate predictors. If they did not contribute significantly, they were removed for parsimony.

For the form recognition outcome, the best-fitting model included main effects of group and item modality, the interaction between group and item modality, and proficiency as the only covariate predictor. PSTM, SRT, and summed TRT did not contribute to model fit, so they were removed from the final model. The best-fitting model (shown in Table 22) for form recognition indicated significant main effects of group ($b = 0.49$, $SE = 0.19$, $p = .012$), item modality for the RO group ($b = 0.57$, $SE = 0.13$, $p < .001$) but not the RWL group ($b = -0.12$, $SE = 0.13$, $p = .35$),

and a group by item modality interaction ($b = -0.69$, $SE = 0.18$, $p < .001$). There was an additional effect of the proficiency covariate on form recognition scores ($b = 0.54$, $SE = 0.12$, $p < .001$). To summarize briefly, participants in the RWL group were significantly more likely to respond to a form recognition item correctly than participants in the RO group, participants in the RO group were significantly more likely to get a correct answer on a form recognition target item in the visual rather than the auditory modality (but not vice versa for RWL), and the group by item modality interaction indicated that the difference in scores between visual and auditory items was greater in the RO group than the RWL group. Additionally, proficiency was a significant predictor of outcomes across participants in both groups for the form recognition task.

Table 22

Best-fitting Model for Form Recognition Outcome

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	0.88	2.41	0.40	1.74-3.35	5.297	<.001	0.49
Group (RWL) ¹	0.49	1.64	0.32	1.12-2.40	2.525	0.012	0.27
Item Modality (Auditory) ¹	0.57	1.76	0.22	1.37-2.26	4.460	<.001	0.31
<i>Covariates</i>							
Proficiency	0.54	1.71	0.21	1.34-2.18	4.318	<.001	0.30
Group (RWL) x Item Modality (Auditory)	-0.69	0.50	0.09	0.35-0.72	-3.780	<.001	-0.382
<i>Random Effects</i>							
1 Participant	Variance	SD					
	0.63	0.80					
1 Item	0.24	0.49					
N _{participant}	119						
N _{item}	25						
Observations	2975						
Marginal R ² / Conditional R ²	0.045 / 0.244						

¹The reading-only group with visual target items is the baseline for this summary.

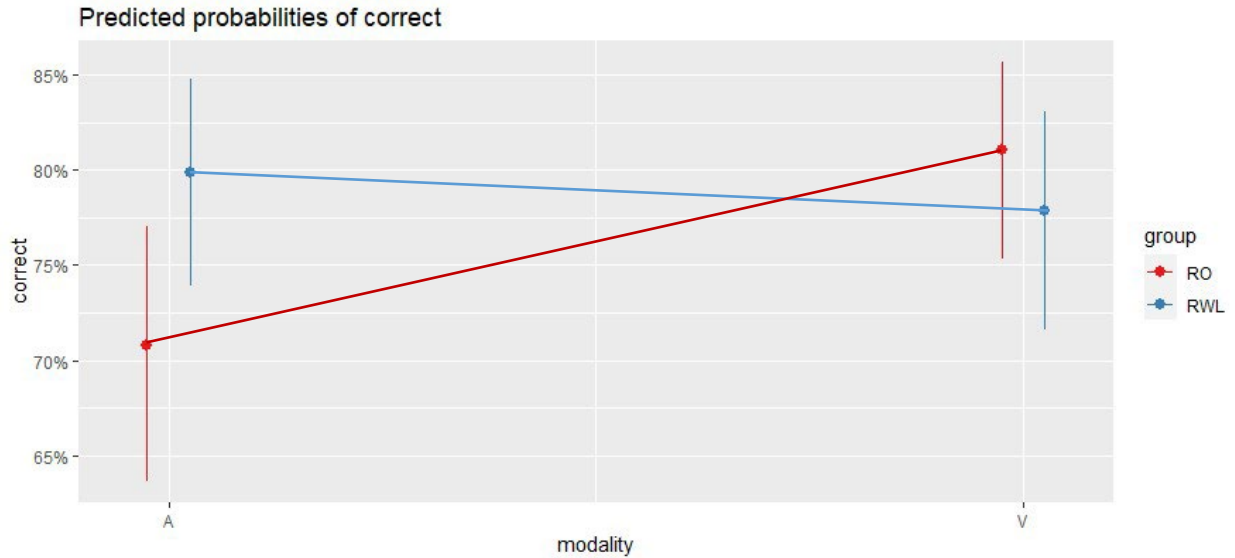


Figure 18. Group by item modality interaction for the form recognition outcome.

For the meaning recognition task, the best-fitting model included main effects of group and item modality, along with the group by item modality interaction and proficiency as a covariate predictor. Again, PSTM, SRT, and summed TRT were not significant predictors during model fitting, so they were removed. Table 23 summarizes results from the best-fitting model for the meaning recognition outcome. The main effect of group did not reach significance ($b = 0.25$, $SE = 0.15$, $p = .088$). As with form recognition, there was a significant main effect in meaning recognition of item modality for RO ($b = 0.30$, $SE = 0.11$, $p < .01$), but not for RWL ($b = 0.12$, $SE = 0.12$, $p = .30$). The model indicated a significant interaction between group and item modality ($b = -0.69$, $SE = 0.09$, $p < .001$), as well as a significant effect of proficiency ($b = 0.49$, $SE = 0.14$, $p < .001$). To summarize, participants in both groups performed similarly in meaning recognition, but participants in the RO group were significantly more accurate on visual than auditory items, but not vice versa in RWL. Again, proficiency was a significant predictor of outcomes.

Table 23*Best-fitting Model for Meaning Recognition Outcome*

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for</i> <i>OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	0.03	1.03	0.20	0.71-1.50	0.161	.872	0.02
Group (RWL) ¹	0.25	1.28	0.19	0.96-1.70	1.704	.088	0.14
Item Modality (Auditory) ¹	0.30	1.35	0.15	1.08-1.69	2.634	<.01	0.17
Group (RWL) x Item Modality (Auditory)	-0.69	0.50	0.09	0.35-0.72	-3.780	<.001	-0.38
<i>Covariates</i>							
Proficiency	0.49	1.63	0.14	1.37-1.92	5.653	<.001	0.27
<i>Random Effects</i>							
1 Participant	0.24	0.49					
1 Item	0.65	0.80					
N _{participant}	119						
N _{item}	25						
Observations	2975						
Marginal R ² / Conditional R ²	0.032 / 0.237						

¹The reading-only group with visual target items is the baseline for this summary.

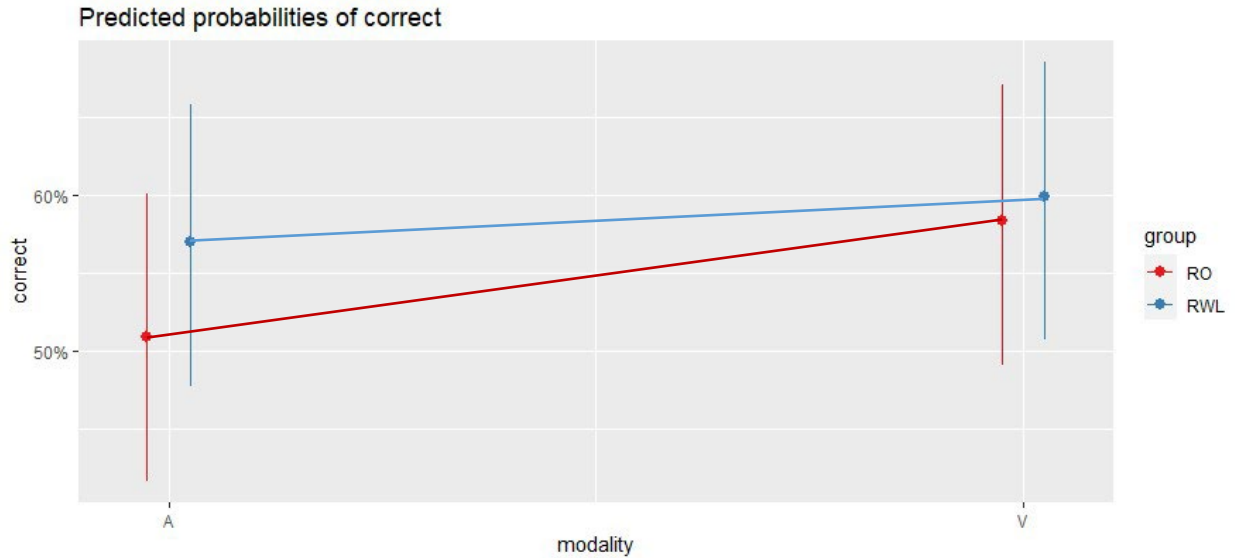


Figure 19. Group by item modality interaction for the meaning recognition outcome.

For the meaning recall outcome, the best-fitting model (summarized in Table 24) included main effects of group and item modality, as well as the group by item modality interaction and proficiency covariate predictor. Again, PSTM, SRT, and summed TRT were not significant predictors, and were removed during model fitting. The best-fitting model indicated significant main effects of group ($b = 1.41, SE = 0.36, p < .001$) and item modality for the RO group ($b = 0.65, SE = 0.21, p < .01$), but not for the RWL group ($b = -0.04, SE = 0.16, p = 0.80$). Additionally, there was a significant group by item modality interaction ($b = -0.69, SE = 0.27, p = .01$), and a significant effect of proficiency ($b = 1.08, SE = 0.22, p < .001$). To summarize, participants in the RWL group outperformed participants in the RO group overall, participants in the RO group performed significantly better on visual than auditory items (but not vice versa for the RWL group), and the group by item modality interaction reflects a significantly wider gap between item performance by modality in RO than in RWL on the meaning recall outcome. Again, proficiency was a significant predictor of performance.

Table 24*Best-fitting Model for Meaning Recall Outcome*

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for</i> <i>OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	-3.7883	0.02	0.01	0.01-0.04	-11.657	<.001	-2.157
Group (RWL) ¹	1.4096	4.09	1.47	2.03-8.27	3.928	<.001	0.78
Item Modality (Auditory) ¹	0.6446	1.91	0.40	1.26-2.88	3.062	.002	0.36
<i>Covariates</i>							
Proficiency	1.075	2.93	0.66	1.89-4.55	4.787	<.001	0.59
Group (RWL) x Item Modality (Auditory)	-0.6859	0.50	0.13	0.30-0.85	-2.575	.01	-0.38
<i>Random Effects</i>							
1 Participant	2.10	1.45					
1 Item	0.47	0.69					
N _{participant}	119						
N _{item}	25						
Observations	2975						
Marginal R ² / Conditional R ²	0.134 / 0.514						

¹The reading-only group with visual target items is the baseline for this summary.

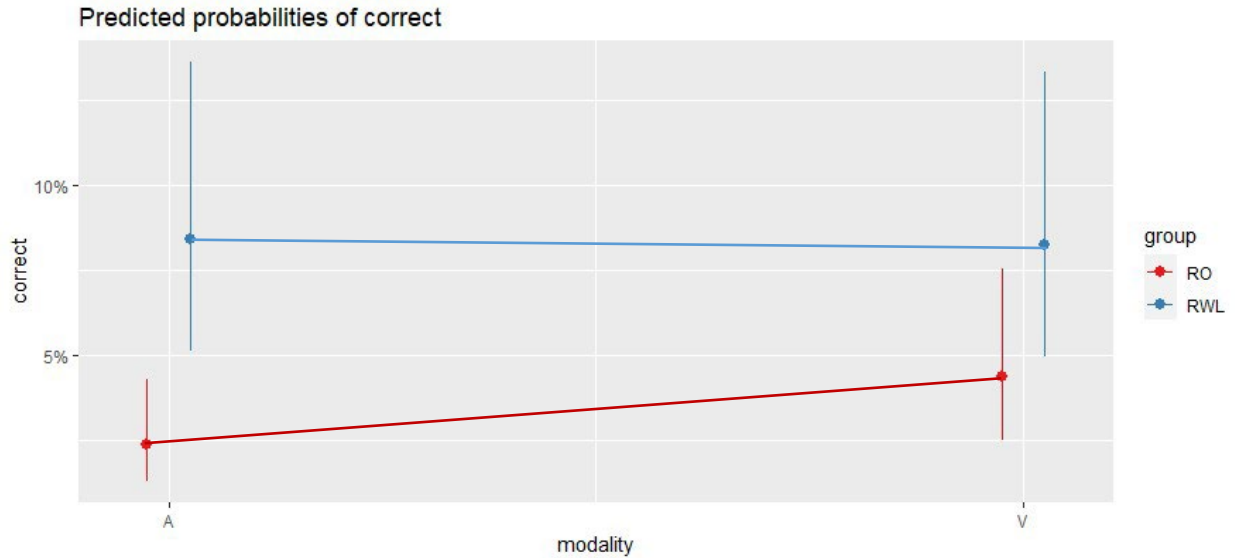


Figure 20. Group by item modality interaction for the meaning recall outcome.

4.6.3 Summary of results for between-group VL outcome comparisons

Table 25 summarizes the main findings of the analyses of the vocabulary learning outcomes, including each research question addressed, predictions, and overall findings for each of the three measures. A very similar pattern emerged across all three tasks regarding group-level differences and within-group differences in test item modality, while proficiency exhibited a robust influence on all three outcomes.

Table 25

Summary of Results from Vocabulary Learning Outcomes

<i>Research question</i>	<i>Learning outcome</i>	<i>Predictions</i>	<i>Findings</i>	<i>Alignment</i>
(1b) To what extent is offline learning of novel word form and meaning different under incidental conditions during reading while listening, compared with reading alone?	Form recognition	*RWL > RO on all posttest outcomes	*Significant effect of group (RWL > RO)	✓
		*Significant effect of item modality in RO (visual > auditory items), but not in RWL	*Significant effect of item modality in RO (visual > auditory items), but not in RWL	✓
		*Significant group x item modality interactions for each of the three outcome measures (in RO, difference between visual and auditory items was greater than in RWL)	*Significant group x item modality interaction (in RO, difference between visual and auditory items was greater than in RWL)	✓
	Meaning recognition	*RWL > RO on all posttest outcomes	*No significant effect of group (RWL = RO, p = .088)	✗
		*Significant effect of item modality in RO (visual > auditory items), but not in RWL	*Significant effect of item modality in RO (visual > auditory items), but not in RWL	✓
		*Significant group x item modality interactions for each of the three outcome measures (in RO, difference between visual and auditory items was greater than in RWL)	*Significant group x item modality interaction (in RO, difference between visual and auditory items was greater than in RWL)	✓

Meaning recall	*RWL > RO on all posttest outcomes	*Significant effect of group (RWL > RO)	✓
	*Significant effect of item modality in RO (visual > auditory items), but not in RWL	*Significant effect of item modality in RO (visual > auditory items), but not in RWL	✓
	*Significant group x item modality interactions for each of the three outcome measures (in RO, difference between visual and auditory items was greater than in RWL)	*Significant group x item modality interaction (in RO, difference between visual and auditory items was greater than in RWL)	✓

4.7 Within-RWL Analyses for Reading Ahead of the Audio

4.7.1 Reading ahead as a predictor of TRT

For the analysis of the effects of reading ahead of the audio on reading time and learning outcomes, the reading-ahead variable was operationalized within the RWL group as the proportion of looking ahead of the audio for the first encounter with each new word in its context, across all individual instances with each word. This created a continuous main predictor of proportion of reading ahead (readahead) for model building, and I specified only the RWL dataset for these analyses. First, a linear mixed-effects model was fitted utilizing the *lmer* function within the *lme4* package in R (Bates et al., 2015), with z-transformed total reading time as the outcome. The best-fitting model is reported in Table 26, with the primary fixed effect as the readehead variable. The analysis revealed that reading ahead of the audio was a significant positive predictor of TRT ($b = 0.09$, $SE = 0.01$, $p < .001$), even when accounting for proficiency. The proficiency covariate had a marginally significant effect on total reading time ($b = -0.17$, $SE = 0.10$, $p = .09$), so it was included in the best-fitting model, as it was considered theoretically important to account for differences in proficiency in assessing the impact of reading ahead of the audio. Additionally, PSTM was a significant covariate predictor of total reading time ($b = 0.25$, $SE = 0.09$, $p < .01$). SRT was not a significant predictor, so it was not included in the best-fitting model.

Table 26*Best-fitting Model for RWL Analysis of Readahead Variable as Predictor of Total Reading Time*

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>t</i>	<i>p</i>	<i>d</i>
Intercept	-0.01	0.09	-0.18 – 0.17	-0.063	0.95	-0.01
Readahead	0.09	0.01	0.07 – 0.12	6.714	<.001	0.20
<i>Covariates</i>						
Proficiency	-0.17	0.10	-0.37 – 0.03	-1.698	.09	-0.45
PSTM	0.25	0.09	0.08 – 0.42	2.841	.005	0.76
<i>Random Effects</i>						
<i>Variance</i>						
1 Participant	0.26					
1 Item	0.27					
N _{participant}	59					
N _{item}	75					
Observations	4425					
Marginal R ² / Conditional R ²	0.047 / 0.572					

4.7.2 Reading ahead as a predictor of learning outcomes

Logistic mixed-effects regression models were fit for within-RWL analyses to each of the three vocabulary learning outcomes (form recognition, meaning recognition, meaning recall). The main fixed effect for each analysis was the readahead variable, with covariate predictors item modality, proficiency, PSTM, SRT, and summed TRT included in the omnibus model. Since TRT was related to reading ahead, it was considered important to account for in the models even if it were not significant within a given individual model of outcomes.

Table 27 indicates the best-fitting model for the form recognition outcome. There was no significant effect of reading ahead of the audio on form recognition outcomes ($b = 0.07$, $SE = 0.09$, $p = .40$). Proficiency was a significant covariate predictor of the form recognition outcome

for the RWL group, as reported earlier in between-group analyses ($b = 0.54$, $SE = 0.18$, $p < .01$), but TRT was not ($b = -0.07$, $SE = 0.08$, $p = .37$). No other covariates (item modality, PSTM, SRT, or TRT) were significant predictors of form recognition scores within the RWL group.

Table 27

Best-fitting Model for RWL Analysis of Readahead Variable on Form Recognition Outcome

<i>Fixed effects</i>	<i>b</i>	<i>OR (exp(b))</i>	<i>SE</i>	<i>95% CI for OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	1.31349	3.72	0.59	2.72 – 5.08	8.267	<.001	0.72
Readahead	0.07014	1.07	0.09	0.91 – 1.26	0.842	0.40	0.04
<i>Covariates</i>							
Proficiency	0.54185	1.72	0.31	1.20 – 2.46	2.958	<.01	0.30
Total reading time	-0.07476	0.93	0.08	0.79 – 1.09	-0.895	0.37	-0.04
<i>Random Effects</i>							
1 Participant	0.64	0.80					
1 Item	0.23	0.48					
N _{participant}	59						
N _{item}	25						
Observations	1475						
Marginal R ² / Conditional R ²	0.04 / 0.24						

Table 28 indicates the best-fitting model for the readahead analysis on the meaning recognition outcome. Again, there was no significant main effect of reading ahead on the meaning recognition outcome ($b = 0.08$, $SE = 0.08$, $p = .312$). Proficiency was a significant covariate predictor of the meaning recognition outcome for the RWL group, as reported earlier in between-group analyses ($b = 0.58$, $SE = 0.13$, $p < .001$), but total reading time was not ($b = 0.002$, $SE = 0.08$, $p = .984$). No other covariates (item modality, PSTM, or SRT) were significant predictors of meaning recognition scores within the RWL group.

Table 28*Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recognition Outcome*

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for</i> <i>OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	0.34977	1.42	0.28	0.97 – 2.08	1.799	0.072	0.19
Readahead	0.07634	1.08	0.08	0.86 – 1.17	1.012	0.312	0.04
<i>Covariates</i>							
Proficiency	0.58029	1.79	0.23	1.38 – 2.31	4.413	<.001	0.32
Total reading time	0.00156	1.00	0.08	0.86 – 1.17	0.02	0.984	0.00
<i>Random Effects</i>							
1 Participant	0.25	0.50					
1 Item	0.75	0.87					
N _{participant}	59						
N _{item}	25						
Observations	1475						
Marginal R ² / Conditional R ²	0.043 / 0.266						

Table 29 summarizes the best-fitting model with the main effect of the readahead variable predicting the meaning recall outcome. There was a significant main effect of reading ahead on the meaning recall outcome ($b = 0.26$, $SE = 0.11$, $p = .018$). Proficiency was a significant covariate predictor of the meaning recall outcome for the RWL group, as reported earlier in between-group analyses ($b = 0.98$, $SE = 0.78$, $p < .001$), but total reading time was not ($b = 0.06$, $SE = 0.11$, $p = .549$). No other covariates (item modality, PSTM, or SRT) were significant predictors of meaning recall within the RWL group.

Table 29*Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recall Outcome*

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for</i> <i>OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	-2.3474	0.10	0.02	0.06 – 0.16	-9.272	<.001	-1.269
Readahead	0.2556	1.29	0.14	1.05 – 1.60	2.368	.018	0.14
<i>Covariates</i>							
Proficiency	0.9784	2.66	0.29	1.50 – 4.73	3.336	<.001	0.54
Total reading time	0.0634	1.07	0.11	0.87 – 1.31	0.599	0.549	0.04
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>					
1 Participant	1.72	1.31					
1 Item	0.39	0.62					
N _{participant}	59						
N _{item}	25						
Observations	1475						
Marginal R ² / Conditional R ²	0.107 / 0.456						

4.7.3 Summary of results for within-RWL analyses

Table 30 summarizes the main findings for the within-RWL analyses to answer RQs 2a and 2b, including the RQ addressed, the outcome of interest, predictions, findings, and alignment of predictions with findings.

Table 30*Summary of Results from Analyses of Reading Ahead of the Audio*

<i>RQ</i>	<i>Outcome</i>	<i>Predictions</i>	<i>Findings</i>	<i>Alignment</i>
(2a) To what extent does reading slightly ahead of the audio in a reading while listening text contribute to learning of novel word form under incidental conditions, as evidenced by faster reading times?	TRT	*Readahead as a significant <i>negative</i> predictor of TRT	*Readahead as a significant <i>positive</i> predictor of TRT	X
(2a) To what extent does reading slightly ahead of the audio in a reading while listening text contribute to learning of novel word form and meaning under incidental conditions, as evidenced by posttest outcomes?	Form Recognition		*No significant effect of readahead variable on form recognition or meaning recognition	X
	Meaning Recognition	*Readahead as a significant predictor of all three learning outcomes		
	Meaning Recall		*Readahead as a significant predictor of meaning recall	✓

4.8 Effects of PSTM on Learning Outcomes

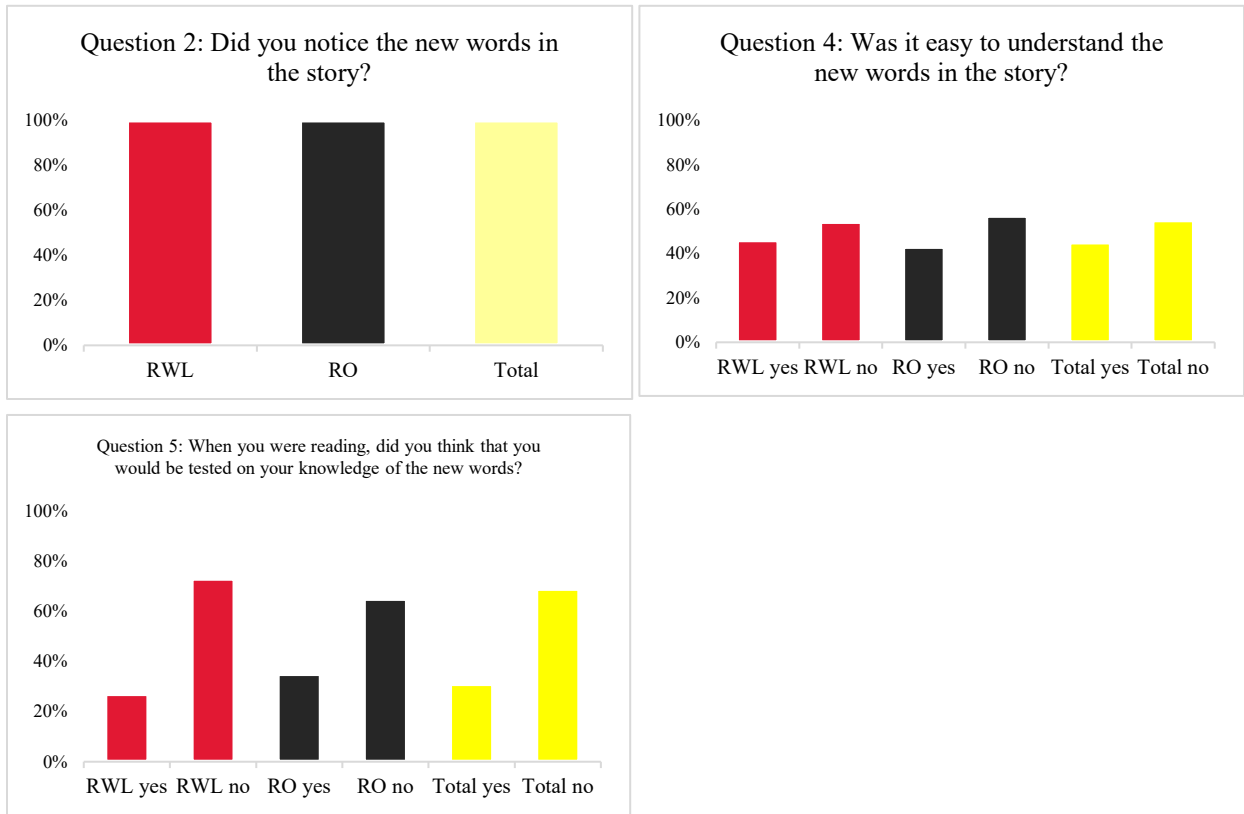
Tables 22-24 above, summarizing the models for each vocabulary learning outcome, provide the information necessary to summarize the results to answer RQ3 regarding the impact of PSTM on learning outcomes. For the participants in this study, given 10 instances of target items in the reading text, there was no impact of PSTM above and beyond the effect of proficiency on vocabulary learning outcomes at both form and meaning levels. As a result, there were also no indications of an aptitude by treatment (ATI) interaction with PSTM and group. Since none of these interactions were detected, they were not graphed.

4.9 Qualitative Debriefing Survey Analysis

Debriefing survey results are summarized in two ways. The dichotomous yes/no answers are summarized in Figures 18-20, while open-ended questions were coded iteratively according to the open coding stage described by Strauss and Corbin (1997), with the goal of reporting emerging themes from the responses. I utilized the answers to the survey questions to code according to these thematic elements. All participants were asked the same six questions on the debriefing survey:

- (1) When you were reading the story, what were you trying to do? (open-ended)
- (2) Did you notice the new words in the story? (dichotomous yes/no)
- (3) If you noticed the new words in the story, did you try to learn them? If so, how?
(both dichotomous and open-ended)
- (4) Was it easy to understand the new words in the story? (dichotomous yes/no)
- (5) When you were reading, did you think you would be tested on your knowledge of the new words?
- (6) Is there anything else about your experience reading and learning the new words you'd like to express that you didn't already express on the previous questions?

Figures 18-20 provide a summary of dichotomous yes/no questions for all participants. All participants ($n = 119$) reported noticing the unfamiliar words in the story, while around half in both groups asserted that it was easy to understand them. Over 2/3 of participants (69% overall; 73% in RWL; 65% in RO) asserted that they did not think they would be tested on the new words, indicating the requisite focus on meaning during the reading task.



Figures 21-23. Group and sample proportions of responses to dichotomous debriefing survey items.

There were several themes that emerged from the iterative open coding process for the open-ended survey items that were given to both groups. For the first survey question (“When you were reading the story, what were you trying to do?”), an overwhelming majority of responses in both groups (85% in RO; 90% in RWL) focused on the theme of *story comprehension*. One RO participant stated, “I was creating an imaginary world where I visualized the story.” Another RWL participant asserted, “I was just trying to absorb the story in my brain as I was supposed to answer some questions based on my understanding of the story.” For the third question (“If you noticed the new words in the story, did you try to learn them? If so, how?”), a single code emerged quickly in responses, focusing on the theme of *contextual clues* in trying to figure out the meaning of the new words. The vast majority of participants both

in the RO group (77%) and the RWL group (88%) made explicit connections in their responses to understanding the meaning of the words from the surrounding context during reading. One participant from the RO group stated, “I tried to make sense of the new words encountered on the basis of how they fit into the context of the sentence and the story as a whole.” Another participant in the RWL group added that they tried to learn “from the context, the words before and after. For example: He drank *...Then I can guess * might be a drink. I also tried to learn from the sentence before and after to guess the meaning.”

There were two additional survey questions on the debriefing survey for participants in the RWL group, that focused specifically on their experiences in reading with simultaneous audio. Two themes emerged from Question 6, which asked “If reading with the audio helped you, in what way(s) did it help?” 22% of the RWL participants included explicit reference to how audio helped in mapping the *pronunciation* of the new words in the story. One participant stated, “it was easier to match the pronunciation of the words with the spelling”. Another reported that the audio “helped me to hear how the new words should be pronounced, and establish the tone.” 39% of the RWL participants responded related to the theme of how the audio assisted in *fluency/pace*. One participant reported that “the audio has a slower pace than my reading speed so it ensured that I can understand the story.” Another asserted that the audio “encouraged me to read the whole sentence in a steady pace.” Another RWL participant reflected on the difference between reading with and without audio: “it made me slow, I can read faster. But when I am reading without an audio, I feel like I forgot something and want to go back and read again. In that regard, the audio helped me not to do that.” It was very interesting to see how many participants (at a very high level of L2 English) noted that their reading was slowed down by the audio, but they saw this as a benefit.

Question 7 asked about the opposite effect of the audio: “If reading with the audio made it more difficult to read, in what way(s) did it make it more difficult?” Unsurprisingly, there were similar mirrored responses to this question as the counter of Question 6, with two theme codes emerging. 54% of RWL participants made comments related to how the audio made reading more difficult in relation to *reading speed/pace*. One participant did not sugarcoat their response: “It makes me read with a slow speed.” Another stated, “I had to make my skimming a little slow because I was trying to align it with the audio.” Another said, “My reading pace was faster and I had to wait for the audio to catch up.” A smaller number of RWL participants (12%) also reported being distracted at times by the audio, with one participant stating, “It is hard to understand the key meaning of the sentences if I read all words by following the audio.” Another reported, “I would like to say, yes, it did also somehow distract me when I was reading the text in a limited time. I felt like I had to follow the audio, or I had to go back to the location the audio was reading if I went ahead too much.”

Finally, question 6 for the RO group and question 8 for the RWL group asked “Is there anything else about your experience reading and learning the new words you’d like to express that you didn’t already express on the previous questions?” A few participants mentioned the challenges of vocabulary learning when the focus of the task was meaning and comprehension, and a few more mentioned that they would have read the new words more and in different ways had they known about the vocabulary tests, but there were no responses that emerged during coding. Given the open-ended nature of the question, this result is unsurprising, and most of the participants expressed their enjoyment of the task.

4.10 Supplemental Analyses

Although planned analyses included creating composite proficiency and PSTM variables, within-construct relationships between the three proficiency measures (cloze, auditory LexTALE, and reading speed) and the two PSTM measures (running memory span, nonword span) were surprisingly weak (see Table 7). As a result, it was determined that exploratory analyses should be completed for each of the research questions, separating out the relative contribution of each measure in the models. These will now be presented in order.

4.10.1 Eye-tracking measures and ID variable supplemental analyses

Tables 31-33 summarize the best-fitting models for between-group analyses of eye-tracking data (GD, TRT, and Visit Count), with all individual difference variables parsed out as individual predictors. The overall analytic structure and all main effects were unaffected by parsing out the individual difference variables. Across the three eye-tracking measures, only reading speed was a significant individual predictor of all three, while the auditory LexTALE measure was a significant predictor of GD and TRT. The cloze measure did not significantly predict any of the eye-tracking processing outcomes. The only significant PSTM predictor among the eye-tracking outcomes was running memory span, which was a significant predictor of visit count.

Table 31*Best-fitting Model for Gaze Duration (ID variables separated)*

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>t</i>	<i>p</i>	<i>Exp(b)</i>
Intercept	5.80	0.02	5.76 – 5.85	244.278	<.001	331.29
Group (RWL) ¹	0.10	0.03	0.05 – 0.15	3.888	<.001	1.11
Instance (Linear)	-0.08	0.01	-0.11 – -0.06	-6.476	<.001	0.92
Group (RWL) x Instance (Linear)	0.04	0.02	0.01 – 0.08	2.517	0.012	1.04
<i>Covariates</i>						
LexTALE	-0.06	0.01	-0.09 - -0.04	-5.026	<.001	0.94
Reading speed	-0.05	0.01	-0.08 - -0.03	-4.007	<.001	0.95
Visit Count	-0.07	0.00	-0.08 - -0.06	-22.279	<.001	0.93
<i>Random Effects</i>						
1 Participant	0.02	0.13				
1 Item	0.01	0.07				
Residual	0.23	0.48				
N _{participant}	119					
N _{item}	25					
Observations	28246					
Marginal R ² / Conditional R ²	0.064 / 0.152					

¹The reading-only group is the baseline for this summary.**Table 32***Best-fitting Model for Total Reading Time (ID variables separated)*

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>t</i>	<i>p</i>	<i>Exp(b)</i>
Intercept	5.73	0.02	5.68 – 5.77	246.832	<.001	307.97
Group (RWL) ¹	0.08	0.02	0.04 – 0.13	3.684	<.001	1.08
Instance (Linear)	-0.20	0.01	-0.23 - -0.17	-14.797	<.001	0.82
Group (RWL) x Instance (Linear)	0.08	0.02	0.05 – 0.12	4.320	<.001	1.08
<i>Covariates</i>						
LexTALE	-0.04	0.01	-0.07 - -0.02	-3.769	<.001	0.96
Reading speed	-0.03	0.01	-0.05 - -0.00	-2.387	0.017	0.97

Visit Count	0.24	0.00	0.23 – 0.24	70.835	<.001	1.27
<hr/>						
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>				
1 Participant	0.01	0.12				
1 Item	0.01	0.08				
Residual	0.26	0.51				
N _{participant}	119					
N _{item}	25					
Observations	28246					
Marginal R ² /	0.167 /					
Conditional R ²	0.227					

¹The reading-only group is the baseline for this summary.

Table 33

Best-fitting Model for Visit Count (ID variables separated)

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>t</i>	<i>p</i>	<i>Exp(b)</i>
Intercept	1.82	0.05	1.73 – 1.92	22.77	<.001	6.17
Group (RWL) ¹	0.93	0.03	0.88 – 0.99	-2.381	0.017	2.53
Instance	0.98	0.00	0.98 – 0.99	-7.771	<.001	2.66
Group (RWL) x Instance	1.01	0.00	1.00 – 1.01	2.152	0.031	2.75
<hr/>						
<i>Covariates</i>						
Reading speed	1.05	0.01	1.03 – 1.08	4.014	<.001	2.86
Running memory span (PSTM)	1.02	0.01	1.00 – 1.05	2.049	0.041	2.77
<hr/>						
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>				
1 Participant	0.01	0.12				
1 Item	0.01	0.08				
N _{participant}	119					
N _{item}	25					
Observations	28246					
Marginal R ² /	0.011 /					
Conditional R ²	0.052					

¹The reading-only group is the baseline for this summary.

4.10.2 Learning outcomes and ID variable supplemental analyses

Tables 34-36 show the results of the best-fitting models for the three learning outcomes in the study, with individual difference variables separated. In terms of main effects, there was no change for either form recognition or meaning recall. For the meaning recognition measure, parsing out the individual difference variables resulted in the group by item modality effect disappearing in the best-fitting model ($b = -0.18$, $SE = 0.14$, $p = 0.265$); however, the within-group differences remained, with participants in RO significantly better at visual than auditory items ($b = 0.30$, $SE = 0.16$, $p = .01$), while no difference existed between visual and auditory item scores within RWL ($b = 0.11$, $SE = 0.12$, $p = 0.33$). Each of the three proficiency measures independently predicted two outcomes (cloze: meaning recognition and meaning recall; LexTALE: form and meaning recognition; Reading speed: form recognition and meaning recall). Conversely, none of the PSTM measures independently predicted outcomes at a significant level, although nonword span as a predictor of form recognition approached significance ($b = 0.27$, $SE = 0.18$, $p = 0.051$). In general, these findings matched the results of the models including composite variables, with proficiency playing a much more prominent role in learning outcomes than memory, and no aptitude by treatment interaction effects detected. One additional deviation from the original analyses came in the finding for meaning recall that total reading time was a significant predictor of learning when the other individual difference variables were parsed ($b = 0.20$, $SE = 0.10$, $p = 0.011$). Total reading time had not been a significant predictor in any of the best-fitting models including the composite proficiency and memory scores.

Table 34*Best-fitting Model for Form Recognition Outcome (ID variables separated)*

<i>Fixed effects</i>	<i>b</i>	<i>OR (exp(b))</i>	<i>SE</i>	<i>95% CI for OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	0.85	2.34	0.38	1.70 – 3.21	5.243	<.001	0.47
Group (RWL) ¹	0.54	1.72	0.32	1.19 – 2.49	2.887	0.004	0.30
Item Modality (Auditory) ¹	0.60	1.82	0.24	1.40 – 2.36	4.531	<.001	0.33
Group (RWL) x Item Modality (Auditory)	-0.69	0.50	0.09	0.35-0.72	-3.769	<.001	-0.38
<i>Covariates</i>							
LexTALE	0.23	1.26	0.11	1.06 – 1.49	2.592	0.01	0.13
Reading speed	0.26	1.30	0.12	1.08 – 1.56	2.76	0.006	0.15
Nonword span (PSTM)	0.27	1.31	0.18	1.00 – 1.72	1.951	0.051	0.15
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>					
1 Participant	0.54	0.73					
1 Item	0.23	0.48					
Modality Participant	0.01	0.001					
N _{participant}	119						
N _{item}	25						
Observations	2975						
Marginal R ² / Conditional R ²	0.05 / 0.246						

¹The reading-only group with visual target items is the baseline for this summary.**Table 35***Best-fitting Model for Meaning Recognition Outcome (ID variables separated)*

<i>Fixed effects</i>	<i>b</i>	<i>OR (exp(b))</i>	<i>SE</i>	<i>95% CI for OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	0.07	1.08	0.21	0.74 – 1.57	0.376	0.707	0.04
Group (RWL) ¹	0.17	1.19	0.18	0.88 – 1.61	1.138	0.255	0.10
Item Modality (Auditory) ¹	0.30	1.35	0.16	1.07 – 1.69	2.566	0.01	0.17

Group (RWL) x Item Modality (Auditory)	-0.18	0.83	0.14	0.60 – 1.15	-1.114	0.265	-0.10
<i>Covariates</i>							
Cloze	0.23	1.26	0.09	1.10 – 1.44	3.271	0.001	0.13
LexTALE	0.20	1.22	0.08	1.07 – 1.40	2.97	0.003	0.11
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>					
1 Participant	0.30	0.55					
1 Item	0.65	0.80					
Modality Participant	0.02	0.003					
N _{participant}	119						
N _{item}	25						
Observations	2975						
Marginal R ² / Conditional R ²	0.034 / 0.239						

¹The reading-only group with visual target items is the baseline for this summary.

Table 36

Best-fitting Model for Meaning Recall Outcome (ID variables separated)

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	-3.73322	0.02	0.01	0.01 – 0.04	-11.716	<.001	-2.157
Group (RWL) ¹	1.35	3.84	1.37	1.91 – 7.72	3.774	<.001	0.74
Item Modality (Auditory) ¹	0.64	1.90	0.40	1.26 – 2.88	3.06	0.002	0.35
Group (RWL) x Item Modality (Auditory)	-0.68	0.51	0.14	0.30 – 0.86	-2.533	0.011	-0.37
<i>Covariates</i>							
Cloze	0.45	1.57	0.26	1.14 – 2.16	2.75	0.006	0.25
Reading speed	0.54	1.71	0.27	1.25 – 2.34	3.39	0.001	0.30
Total reading time	0.20	1.22	0.10	1.04 – 1.42	2.433	0.011	0.11
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>					
1 Participant	1.94	1.39					
1 Item	0.46	0.68					
N _{participant}	119						
N _{item}	25						

Observations	2975
Marginal R ² /	0.139 /
Conditional R ²	0.503

¹The reading-only group with visual target items is the baseline for this summary.

4.10.3 Reading ahead of the audio in RWL and ID variable supplemental analyses

Tables 37-40 summarize the best-fitting models focused on reading ahead of the audio within RWL as a predictor of TRT and learning outcomes. There were no differences in main effects of reading ahead of the audio from the analyses separating out individual difference variables, with reading ahead remaining a significant positive predictor within RWL of TRT and the meaning recall learning outcome. The LexTALE proficiency measure was a significant covariate predictor both of form and meaning recognition, while reading speed was a significant predictor of meaning recognition and recall in the RWL group. The cloze proficiency measure did not significantly predict any of the learning outcomes. No PSTM measures were significant covariate predictors in any of the models of the RWL group.

Table 37

Best-fitting Model for RWL Analysis of Readahead Variable as Predictor of Total Reading Time (ID variables separated)

<i>Fixed effects</i>	<i>b</i>	<i>SE</i>	<i>95% CI</i>	<i>t</i>	<i>p</i>	<i>d</i>
Intercept	-0.01	0.09	-0.19 – 0.17	-0.077	0.939	-0.006
Readahead	0.09	0.01	0.07 – 0.12	6.660	<.001	0.05
<i>Covariates</i>						
Serial reaction time	-0.13	0.07	-0.27 – 0.01	-1.847	0.065	-0.07
Nonword span (PSTM)	0.10	0.06	-0.02 – 0.23	1.613	0.11	0.05
<i>Random Effects</i>	<i>Variance</i>					
1 Participant	0.27					

1 Item	0.27
N _{participant}	59
N _{item}	75
Observations	4425
Marginal R ² / Conditional R ²	0.046 / 0.574

Table 38

Best-fitting Model for RWL Analysis of Readahead Variable on Form Recognition Outcome (ID variables separated)

<i>Fixed effects</i>	<i>b</i>	<i>OR (exp(b))</i>	<i>SE</i>	<i>95% CI for OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	1.31	3.69	0.60	2.69 – 5.07	8.053	<.001	0.72
Readahead	0.09	1.09	0.09	0.93 – 1.29	1.085	0.278	0.05
<i>Covariates</i>							
LexTALE	0.28	1.33	0.17	1.04 – 1.70	2.258	0.024	0.16
<i>Random Effects</i>							
1 Participant	0.70	0.83					
1 Item	0.23	0.48					
N _{participant}	59						
N _{item}	25						
Observations	1475						
Marginal R ² / Conditional R ²	0.04 / 0.24						

Table 39*Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recognition Outcome**(ID variables separated)*

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for</i> <i>OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	0.39	1.48	0.29	1.01 – 2.17	1.989	0.047	0.22
Readahead	0.08	1.09	0.08	0.94 – 1.26	1.119	0.263	0.05
<i>Covariates</i>							
LexTALE	0.29	1.34	0.12	1.13 – 1.59	3.302	0.001	0.16
Reading speed	0.26	1.30	0.14	1.05 – 1.61	2.433	0.015	0.15
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>					
1 Participant	0.26	0.50					
1 Item	0.75	0.87					
N _{participant}	59						
N _{item}	25						
Observations	1475						
Marginal R ² / Conditional R ²	0.042 / 0.266						

Table 40*Best-fitting Model for RWL Analysis of Readahead Variable on Meaning Recall Outcome (ID**variables separated)*

<i>Fixed effects</i>	<i>b</i>	<i>OR</i> <i>(exp(b))</i>	<i>SE</i>	<i>95% CI for</i> <i>OR</i>	<i>z</i>	<i>p</i>	<i>d</i>
Intercept	-2.20	0.11	0.03	0.07 – 0.18	-9.166	<.001	-1.22
Readahead	0.27	1.31	0.14	1.07 – 1.61	2.582	0.01	0.15
<i>Covariates</i>							
Reading speed	0.82	2.27	0.50	1.47 – 3.49	3.724	<.001	0.45
<i>Random Effects</i>	<i>Variance</i>	<i>SD</i>					
1 Participant	1.61	1.27					
1 Item	0.39	0.62					
N _{participant}	59						

N _{item}	25
Observations	1475
Marginal R ² /	0.114 /
Conditional R ²	0.448

4.11 General Summary of Results

In this chapter, I have laid out the results of the study in the order of the research questions. I summarized the evidence of a focus on meaning for participants through the high scores on the reading comprehension questions, and then summarized both descriptive and inferential analyses for eye-tracking patterns (for GD, TRT, and visit count, see Table 18), vocabulary learning outcomes (for form recognition, meaning recognition, meaning recall, see Table 25), and reported analyses of reading ahead as a predictor of TRT and learning outcomes (see Table 30). Additionally, I reported the results of PSTM as a (non-)predictor of vocabulary learning for both groups of participants, and described and reported the coding process and results from the qualitative survey data analysis. Based on the weak relationships between individual difference variables, I conducted and reported separate analyses parsing out the relative effects of these individual measures on processing and learning outcomes. We now turn to the final chapter to discuss these findings.

Chapter 5: Discussion

5.1 Introduction to Discussion

In the final chapter, the findings of the present study are discussed systematically, in order of the research questions, and in light of initial predictions from the dissertation proposal. Findings are connected to other studies, with broader theoretical implications on how the current findings inform general understanding of L2 lexical development. The reported experiences of the participants in the study are discussed, along with consideration of how those experiences may have influenced outcomes. Various limitations of the study are summarized, along with pedagogical and research design implications, along with a number of future directions that can expand on the novel methods utilized and findings revealed here. Table 41 provides a summary of the overall predictions from the dissertation proposal, and how the results from the present study align or diverge with predicted outcomes.

5.2 Multimodal Reading and L2 Lexical Development

5.2.1 The process of reading ± audio: stability, trajectory, and change over time

5.2.1.1 Early form familiarity development: gaze duration

GD, or the total time spent looking at the area including the target word during the first encounter with it on a page, has been utilized frequently in L2 eye-tracking studies as an index of form familiarity (see Conklin et al., 2018; Godfroid, 2020). It is thought to capture the phenomenon described by the EZ-Reader model (e.g., Pollatsek et al., 2006) as the *familiarity check*, wherein the gaze of the reader is focused very briefly on a word to verify form prior to automatic programming eye movement to the next word during fluent reading. Models such as

Table 41

Research Questions, Predictions, and Reported Findings in the Present Study

RQ	Prediction(s)	Comparison	Initial predictions		Findings	Alignment
			<u>Online Eye tracking</u>	<u>Offline measure</u>		
1a	1-2	Between-group	<ul style="list-style-type: none"> Decrease in reading times across 10 instances for both groups 		<ul style="list-style-type: none"> Significant decrease in all three ET measures for both groups 	✓
			<ul style="list-style-type: none"> RO > RWL in GD and TRT, evidenced by significant group difference in GCA analyses 		<ul style="list-style-type: none"> RWL > RO in GD and TRT 	✗
			<ul style="list-style-type: none"> RO > RWL in total visit counts to target items 		<ul style="list-style-type: none"> RO > RWL in visit counts 	✓
1b	3-5	Between-group		<ul style="list-style-type: none"> RWL > RO on all posttest outcomes in both item modalities 	<ul style="list-style-type: none"> RWL > RO overall in form recognition and meaning recall; RWL = RO in meaning recognition (p = .088) 	Partial
				<ul style="list-style-type: none"> Larger effects in auditory items 	<ul style="list-style-type: none"> Large effects of item modality in RO, but not RWL 	✓
				<ul style="list-style-type: none"> Group x item modality interaction (RO visual items > RO auditory items, but not vice versa)¹ 	<ul style="list-style-type: none"> Group x item modality interaction (RO visual items > RO auditory items, but not vice versa) 	✓
2a	6	Within-RWL	<ul style="list-style-type: none"> Proportion of reading ahead as a significant <i>negative</i> predictor of TRT in RWL 		<ul style="list-style-type: none"> Proportion of reading ahead as a significant <i>positive</i> predictor of TRT in RWL 	✗

2b	7-9	Within-RWL	<ul style="list-style-type: none"> Proportion of reading ahead as a significant predictor of form recognition, meaning recognition, and meaning recall posttest outcomes 	<ul style="list-style-type: none"> Proportion of reading ahead did not predict form or meaning recognition outcomes Proportion of reading ahead as a predictor of meaning recall outcome 	X
3	10	Between-group	<ul style="list-style-type: none"> PSTM as a predictor of FR and MR outcomes, interacting with group 	<ul style="list-style-type: none"> PSTM did not predict learning outcomes on any of the three measures; no aptitude by treatment interaction 	X

¹This interaction effect was found for meaning recognition only in models including composite individual difference variables, but the effects of item modality indicated that participants in RO were superior in visual items compared with auditory items, but not vice versa in RWL.

EZ-Reader are based on assumptions that processing in L1 reading mimics auditory processing in being serial in nature, necessarily producing sublexical or prelexical spoken representations in memory (see Pollatsek et al., p. 39), and include the familiarity check as an essential preliminary stage of processing. Cop et al. (2015) extended the argument to L2 reading, arguing that the familiarity check functions as a “trigger” for subsequent oculomotor programming of a saccade toward the following word (p. 3). In cases where familiarity with a word does not exist, as in encountering new vocabulary, the familiarity check should result in slower initial reading times, particularly in polysyllabic words, with a gradual speed-up across exposures (see Godfroid, 2020b). Interruptions to form familiarity and subsequent lexical access would also induce more revisiting of the interest area with the targets, especially in cases where meaning comprehension is impacted. These interruptions have been found to be even more evident during the familiarity check in L2 reading (Cop et al., 2015).

GD was utilized in the present study to examine the development of the familiarity check across encounters, in order to answer the first research question regarding the developmental trajectory of form acquisition of new words. While both RO and RWL group GD decreased significantly from 1-10 instances, participants in the RWL group exhibited consistently *longer* GD times across the ten exposures to new words, counter to my initial predictions for faster form familiarity from RWL. I attribute this to two possible reasons. First, the nature of the task meant that RWL participants were required to have a slower reading process in following along with the audio recording, progressing much more slowly than RWL participants could read overall (audio rate: 160-180 WPM; Mean reading speed for RWL = 225 WPM; SD = 62.8). Even though reading speed was accounted for in the comparative models, it appears that the nature of the RWL task obliged participants to read more slowly, especially given that they were explicitly

instructed and reminded to read along with the audio. Along with the more stable late measures of reading time and visit counts, the overall picture of reading new words was different in RWL than in RO. When reading with the audio, participants in this study spent more initial time looking at new pseudowords, exemplified in significantly longer GD times. This additional attention, coupled with the simultaneous phonological exemplars provided by the audio, likely assisted participants in form-form encoding during initial exposures, potentially freeing up proportionally longer GD time during later instances for higher-level meaning integration processes (which were revealed through better scores on learning outcome measures). Since the familiarity check/GD times sped up at a more leisurely pace in the RWL group (shown in the group by instance interaction in the GD models), participants in RWL appear both to have had more time and done more with it in fixating on unfamiliar words.

A second related reason connects to the argument I made in Malone (2018), that it is highly possible that the presence of audio in the input facilitated attention and focus on the new words and deeper processing in a bidirectional manner of sound and symbol in mapping phonology to orthography in new L2 vocabulary. In this account, the provision of audio during earlier instances would tap into individual differences in PSTM resources during the initial establishment of form-form connections (as I found in Malone, 2018), but after 10 encounters would be revealed only through better form-meaning connections (as reflected in outcomes in this study). These findings regarding longer reading times provide additional evidence for an account for benefits of RWL that is grounded in deeper processing and more cognitive resources freed up for processing new vocabulary during RWL than RO.

In some ways, this finding related to GD is slightly counterintuitive to traditional models of L2 lexical access and reading in naturalistic environments across multiple repetitions with

new words. These studies have typically argued for the position that reducing speed is the goal in solidifying lexical access during L2 reading, and maximizing reading fluency (see the discussion in Pellicer-Sanchez, 2016). However, the findings from Elgort et al. (2018) indicate that L2 reading times for new words remain slower than for familiar words even across many exposures. Therefore, the slightly slower rate of processing for RWL reported here provides initial evidence that developing lexical familiarity may be nuanced in additional ways by the multimodal input. While the RWL reading process mirrored RO in a decrease of time across encounters, it remained longer, but likely more productive. As I will discuss below, this may be tied to the specific instructions to participants in the RWL group, but it reflected unexpected processes, and did not fit with my initial prediction from the study proposal.

This finding aligns in some ways with other recent findings in L2 RWL studies (e.g., Conklin et al., 2020; Tuzcu, 2023). Especially during early encounters with new words, Long (2017) argued that the presence of auditory input could increase perceptual salience of the targets, over and above RO, drawing additional attention to the forms especially during the first few encounters with them in the text. This could account for the result of longer GD times in RWL, particularly among higher-proficiency L2 readers, which was found with the pseudoword targets in the present study. Tuzcu (2023) found that GD times demonstrated a significant decrease in the RWL group only after the 10th exposure to novel pseudowords, whereas there was a significant decrease in the RO group during the first 10 instances in the text, largely mirroring the findings in GD in the current study. Here, there was a significant decrease in GD from the 1st to 10th instance in RWL, but the same general pattern emerged through the group by instance interaction.

This also may indicate that RWL during early encounters can serve as an attention-directing process, focusing the cognitive spotlight (i.e., eye fixations) on unfamiliar words in context more slowly than in RO, increasing their salience and the likelihood of detection, noticing, and awareness in the input (Long, 2017). However, the account within noticing would also predict that the rate of GD would become faster more quickly across exposures in RWL, given that the first few instances serve as explicit and noticed exemplars, which runs counter to the findings here of a group by instance interaction, with RO reducing in GD more quickly than RWL. As discussed below, it is more likely that RWL participants were intentional in aligning their reading with the slower audio, which resulted in slower GD across the 10 instances. Regardless of how they were noticed, more stable and longer initial encounters with new words would also result in stronger and more consistent form-form connections, which would enable deeper semantic processing during subsequent encounters.

Interestingly, this obligatory deceleration of initial time during the familiarity check during the first 10 encounters with new words in RWL compared to RO appears to have accelerated acquisition and learning, as form-form familiarity was more fully developed and semantic outcomes revealed better learning from RWL (as I will discuss below). Slower initial GD, coupled with the likelihood of direct mapping of phonology to orthography during reading with the presence of audio exemplars of new word pronunciations, provides clear evidence of online processing benefits of RWL, albeit in a counterintuitive way in terms of speed. In my initial predictions, I did not account for this obligatory slowdown, assuming that participants would probably ignore the instructions and speed up more quickly through the reading in RWL across time, as form-form connections were established. That may have been the case at early

exposures, but sustained longer GD in subsequent encounters (and longer TRT across later instances) may have contributed to deeper processing.

While there was some evidence of speeding up GD times overall within RWL (as evidenced by the significant effect of instance in the model for both groups), the longer initial reading at each instance, coupled with the interaction between group and instance, indicated that these effects were important and long-lasting to the reading process. As such, the faster speed-up in GD may come at a semantic cost for RO in new word learning, compared with RWL, and particularly in a task with focus on comprehension of meaning. This finding could have very interesting possible pedagogical implications (discussed below), but will need to be further verified across proficiency levels to begin to address the optimal ratio of reading to audio rate at the individual level.

This finding regarding the initial familiarity check of a longitudinal decrease in gaze duration within the first 10 exposures within RWL in L2 reading aligns with Tuzcu (2023), who found a similar pattern of longer gaze durations in RWL, and a significant decrease in gaze duration in the RWL group. However, Tuzcu (2023) also reported no significant speed-up in gaze duration in the RO group. This difference of findings may be attributed to the fact that in Tuzcu (2023), participants in the RO group were not under time pressure for the task, but instead told to press a button when they finished reading each trial to progress to the next one. Participants in the present study were aware that the reading time was limited, which may account for the significant decrease in GD across encounters in RO, as participants were more urgently trying to read for meaning.

One final note on the GD results as it corresponds to the familiarity check is at the theoretical level. Recent work in reading research has reported mixed findings regarding the

possibility of parafoveal processing (1-5 degrees away from the area of fixation) as a way of previewing upcoming text. While there is some evidence for orthographic information to be previewed in parafoveal vision, at least in L1 reading (see Antúnez et al., 2021), there is not consensus on whether phonological and/or semantic information are accessed. However, the obligatory slowdown for L2 reading in the present study may function in an analogous way to parafoveal effects as a preview of information about new words. Given the slower rate of audio than comfortable reading speed for the participants, along with explicit instructions to read along with the audio, participants would be reading more slowly than in RO. As discussed below, in cases where the eye fixations were slightly ahead of the audio, eye fixations in foveal vision during the familiarity check would be more likely to pick up on unfamiliar forms, acting as a warning sign that would then be almost immediately confirmed with the auditory exemplar. The visual preview provided by reading ahead, as discussed below, may function as a benefit.

5.2.1.2 Late form and meaning integration: TRT and visit count

In general, cognitive control models of reading define *total reading time* (TRT) as a late measure of word-to-text integration during reading (for L1, see Pollatsek et al., 2006; for L2, see Godfroid, 2020). Other studies of L2 vocabulary acquisition while reading have found a consistent pattern of a decreasing trajectory in TRT across exposures to new words, whether real rare words or pseudowords (Pellicer-Sanchez, 2016; Godfroid et al., 2018; Mohamed, 2018; Elgort et al., 2018). These studies have reported similar S-shaped patterns to lexical development from RO regardless of word type, which mirror vocabulary acquisition from reading in the L1. To my knowledge, only one (Tuzcu, 2023) has made a direct comparison between RO and RWL reading patterns in a naturalistic reading task, and found a similar S-shaped curvilinear decrease across exposures in both groups. However, Tuzcu (2023) did not include specific task

instructions to read along with the audio, nor reminders to do so. I included both in my study, along with monitoring of audiovisual alignment by the attending researcher. This likely contributed to a smoother line and pattern of TRT across instances of new words in the RWL group in the present study. Since participants were focused on aligning their reading with the audio, the decrease in TRT was steady as form familiarity progressed, but less pronounced than in the RO group or in Tuzcu (2023). Again, a reasonable explanation for the divergence in both GD and TRT trajectory across exposures (fitted best by a *linear* regression term, rather than cubic or quadratic), both from Tuzcu (2023) and my own predictions from the dissertation proposal, is that participants in the present study were doing a better job of reading along with the audio, which was considerably slower than their reading speed. Although pilot testing for the current study had resulted in the speed of the audio being increased for the final study, the participant demographic for the final study included a noticeably higher level of L2 English proficiency than for the pilot. This means that the rate of slowdown of reading due to the audio was likely even more noticeable. Faithfully following the task instructions appears to have resulted in direct impact on the reading *process* for new words, and a positive impact on the reading *product* (new word learning) that did not emerge in Tuzcu (2023).

As discussed earlier, after a review of the literature on *regression count* as a late measure of reading stability and semantic processing, I concluded that a better eye-tracking measure of repetitions/exposures to novel words is *visit count* (see Godfroid, 2020b for a description). Given the proliferation of research into frequency of exposure as an important component of new word learning from context in the L2 (e.g., Hulstijn et al., 1996; Horst et al., 1998; Waring & Takaki, 2003; Webb & Chang, 2015), and recent work on the role of statistical learning in L2 more broadly, especially during reading (Ren & Wang, 2023), I considered visit count to be a better

metric of total encounters with the new words for participants. As such, it was included in the analyses.

The RWL group exhibited a more stable and consistent reading process, measured both in TRT and visit count, than the RO group, even as reading times and visit count decreased more sharply over time in the RO group. Participants in this study in the RO group appear to have become familiar with the new word forms in the text, especially around instance 6-7, following findings from several other recent studies (Pellicer-Sanchez, 2016; Godfroid et al., 2018; Mohamed, 2018; Elgort et al., 2018; Tuzcu, 2023). As this familiarity grew, participants in the RO group decreased in total visits to targets, more closely matching the stable pattern of reading in the RWL group. When the total visits decreased in RO, differences in TRT were revealed, in that the RO group was reading the pseudowords for less TRT than RO, better aligning with observed differences in GD patterns after instance 6 (see Figures 3-5 above for visuals).

5.2.1.3 Summary of between-groups eye-tracking discussion

These findings regarding reading patterns in RWL and RO replicate Tuzcu (2023) in the difference in GD and TRT in RWL compared to RO, and the work of others in a gradual decrease in TRT across exposures to new words while reading in an L2. To my knowledge, this finding of a difference in reading patterns revealed by visit count data has not been previously discovered. The present study clearly replicated other findings regarding the developmental trajectory of form familiarity and online processing of new words in RO, while both aligning with and slightly diverging from recent findings regarding similarities and differences in RWL. Participants in the RWL group exhibited longer gaze durations, fewer total visits to target items, and a smoother but steady decline in GD and TRT across the 10 instances of each pseudoword in the text of the story. Importantly, the finding regarding longer GD indicates that the initial

familiarity check for these participants was decelerated by the presence of audio in the input, making initial encounters with new words necessarily longer in nature, more likely to attract attention, and likely more productive in mapping form-form connections and facilitating subsequent processing for meaning (Long, 2017). Additionally, and crucially, while participants in RO visited the target words more in total, it did not contribute to better learning outcomes.

It should be restated that these findings for GD and TRT differ from my initial predictions, where I asserted that the developmental trajectory would indicate a faster decrease in both GD and TRT for the RWL group. In fact, I found the opposite. However, the results from this study align closely with other recent pioneering work in comparing reading patterns in RWL and RO during L2 reading (Tuzcu, 2023), and make logical sense when considering the clear impact of task instructions and audio rate on the reading process. Given recent evidence from Tuzcu (2023) that longer reading times are connected with explicit awareness of new words in the text, along with theoretical support for the relationship between reading times and the learning of word meanings (e.g., Godfroid et al., 2018; Mohamed, 2018), the explicit instructions given to participants in the present study to read along with the audio appears to have facilitated longer initial and total reading times, more noticing, fewer returns to the targets, and more subsequent time allocated to meaning integration, and thereby greater explicit learning gains from RWL.

This account for the relationship between attention, noticing, and explicit awareness during new word learning in RWL also indicates that the process for learning new words during reading in the present study was likely explicit in nature, even under time pressure and utilizing strictly-defined incidental conditions, at least for participants at high levels of L2 proficiency. The clear evidence from the qualitative survey data also indicated that participants were aware of

the unfamiliar words in the text, and many indicated strategies they employed in trying to figure out their meanings, even while under time pressure. The additional reading times for GD and TRT for the RWL group almost certainly influenced awareness, especially during the first few encounters with the pseudowords in the story.

5.2.2 The products of reading ± audio

5.2.2.1 Overall learning gains of RWL over RO

To answer RQ1b regarding the relative benefits of RWL on vocabulary learning, I utilized three learning tasks (form recognition, meaning recognition, meaning recall), roughly corresponding to three progressive stages of word familiarity and lexical development (see Webb & Nation, 2017 for a discussion), to examine differential effects of treatment group (RO vs. RWL) on pseudoword learning outcomes. Significant group effects across all items were found for form recognition and meaning recall (RWL > RO), but not meaning recognition, and summed TRT was not a significant predictor of learning outcomes in either group, when a battery of other cognitive individual differences were accounted for in the models. The lack of an effect of summed TRT as a predictor of form recognition replicates other recent findings in comparable L2 studies (Pellicer-Sanchez, 2016; Godfroid et al., 2018; Tuzcu, 2023), although two (Pellicer-Sanchez, 2016; Godfroid et al., 2018) reported TRT as a significant predictor either of meaning recognition or meaning recall. However, these studies also included far fewer measured individual difference variables, or reported only self-rated L2 proficiency. Accounting for a more robust set of cognitive individual difference variables may explain the lack of effects of summed TRT on either form or meaning outcomes, for participants at a very advanced level of L2 proficiency.

Several other recent studies (e.g., Malone, 2018; Chen, 2021) have reported benefits of RWL on both form and meaning learning of new words, although two (Brown et al., 2008; Tuzcu, 2023) did not find that superior scores reached statistical significance for meaning recognition gains, matching the results of the present study. The finding of no significant effect of group ($p = .088$) on the meaning recognition outcome in the present study is surprising. However, potential effects could be obscured by the nature of the multiple-choice outcome measure. In this task, I included only three total answer choices, with the correct answer and two distractors. Including additional distractors may have revealed group differences in meaning recognition across all items, rather than just auditory items. Additionally, other studies reporting group effects of RO versus RWL in favor of RWL on meaning recognition outcomes (Malone, 2018; Chen, 2021) measured learning from only two or four exposures to target items. Especially given the high level of L2 proficiency for participants in the present study, the additional encounters with the new words may have resulted in comparable overall learning from RWL and RO, especially on visual items, in the easier semantic task (meaning recognition). Given the fact that the significant group difference in form recognition was due exclusively to the large effect of item modality and a group by modality interaction, it is not particularly surprising that group differences did not emerge in the meaning recognition task. In this line of reasoning, differences in the knowledge of meaning would only be revealed through the auditory items on the meaning recognition task (which were evident), and the more difficult meaning recall task.

Even without summed TRT as an individual-level predictor of learning outcomes in either group, the significant group-level differences in reading times likely influenced group-level form and meaning learning outcomes, especially in meaning recall. The twin findings of greater time on task and superior learning outcomes in RWL indicates a strong possibility that

when even high-proficiency L2 readers are provided with auditory input, stronger form-form connections are established through an enforced deceleration of the familiarity check and longer TRT during reading. The evidence from the present study that deeper semantic processing occurred from RWL, as shown through superior meaning recall scores both on visual and auditory items, supports this hypothesis of longer and more stable initial reading times as evidence within RWL for benefits. Given the way these findings of group-level differences in learning outcomes align with several recent studies of new word learning (e.g., Malone, 2018; Teng, 2018; Chen, 2021), but not the most comparable recent study including a meaning recall measure (Tuzcu, 2023), additional research is urgently needed, especially in a more direct comparison of task instructions within RWL and treatment-test item modality congruence.

From a theoretical perspective, across prominent models both of L1 and L2 word recognition (e.g., Coltheart, 2005; Perfetti, 2007), phonological and orthographic information are both essential aspects of recognition and integration of information in effective reading (see Rayner et al., 2012). As such, providing both in RWL appears to have facilitated stronger form-form connections during the initial development of lexical representations of the pseudowords in the present study, which then allowed for deeper semantic processing across 10 instances of the pseudowords in the story. These findings suggesting better form encoding and deeper semantic processing from RWL also provide some evidence for a more nuanced version of Paivio's influential Dual Coding Theory (Paivio & Csapo, 1973). During phonological and orthographic mapping of new words in an L2, it appears that multiple simultaneous verbal streams of information (orthographic/visual + phonological/auditory) may provide additive effects over visual information alone. The findings also extend support for a continued connection in linguistic contexts to a Redundant Signals Effect for L2 vocabulary learning (Kinchla, 1974;

Lewandowski & Kobus, 1993; Montali & Lewandowski, 1996). At least for participants at a high level of L2 proficiency, there was no indication of cognitive overload (e.g., Sweller, 1987); participants in the RWL group responded accurately to comprehension questions, learned equally well or better on all visual outcome items, and performed significantly better on auditory outcome items across the three tasks.

5.2.2.2 Phonology and equitable test item modality

One of the unique contributions of the present study was the randomized presentation of test items across both visual and auditory modalities in the outcome measures. This design allowed for direct measurement of the benefits of phonological mapping that simultaneous audio provided in RWL. Including both modalities in test items followed calls from multiple sources within L2 lexical development regarding the importance of congruence between treatment and test item modality (e.g., Peters & Montero Perez, 2015; Hatami, 2017; Jelani & Boers, 2018; Uchihara et al., 2022), including the assertion that visual-only test items can bias findings toward visual modalities for group-level comparisons in multimodal research. This bias within traditional designs under incidental conditions was clearly revealed in the present study, as participants in the RO group performed significantly better across all three learning outcomes on visual than auditory items, while participants in the RWL group performed equally well on both item modality types on all three tasks. This study provides important evidence not only that test item modality can obscure learning gains from multimodal conditions (which might explain Tuzcu's (2023) findings of no group-level learning differences), but also compelling evidence that the deceleration of the familiarity check during initial encounters with new words (as evidenced by significantly longer gaze duration in RWL), coupled with auditory exemplars of pseudoword pronunciation alongside orthographic information, appears to result in accelerated

phonological form mapping. To my knowledge, no other recent study within Applied Linguistics/SLA or in other domains has utilized multiple test item modalities to measure learning gains in vocabulary learning from RWL versus RO under incidental conditions.

The inclusion of auditory test items allowed for group differences in phonological aspects of developing word knowledge to emerge across the three tasks, which otherwise would not have been evident. Tuzcu (2023) found no evidence of learning gains in RWL vs. RO groups among similar L2 English learners reading a very similar-length story, but did not account for test item modality as a variable. Sublexical phonological processing has been shown to be prominent during reading new words in English (e.g., Coltheart, 2005; Rayner et al., 2012), and there is growing consensus that phonological information is automatically accessed in reading across languages and script types, both in L1 and L2 (e.g., Wang et al., 2003; Brysbaert, 2022; Zhang et al., 2023). Coupled with the tight theoretical link between orthography, phonology, and semantic processing across the development of L2 lexical representations (Jiang, 2000; Bordag et al., 2021), it is no surprise that participants in the present study were helped in learning phonological forms (assessed through auditory outcome items) by the presence of phonological exemplars during RWL. These findings provide additional evidence for the benefits of RWL in accelerating phonological development in new word representations from the presence of audio in the input.

One of the most consistent findings of the study (in line with predictions) was the treatment by test item modality interactions in each of the learning outcomes. For each outcome, participants in the RO group were significantly better in visual than auditory items, but participants in the RWL group performed equally well on both item modality types. This is perhaps the clearest evidence in the present study of phonological benefits of RWL on learning outcomes without impacting scores on visual items, and is notable in being the first study to

uncover this interaction. While there was an initial speed cost in terms of reading rate, the reading pattern of RWL matched other studies in its downward trajectory, and when total time on task was controlled, the phonological benefits of RWL were striking. Given theoretical accounts that universally assume that the reading process in English and many other L2s involves phonology (e.g., Pollatsek et al., 2006; Brysbaert, 2022), along with the centrality of phonology to lexical development in the L2 (Jiang, 2000; Gor et al., 2021; Bordag et al., 2022), it comes as little surprise that there are facilitative effects in learning new words when simultaneous phonological support is provided.

Crucially, the learning benefits of simultaneous audio did not come at the expense of orthographic learning, as scores on visual items were equal between groups on the form and meaning recognition tasks, and superior from RWL on the meaning recall outcome. The superior effects of RWL on auditory items also uncovered group-level differences in form and meaning recognition outcomes that would not have been revealed by visual-only items. This could help explain some recent findings (e.g., Brown et al., 2008; Tuzcu, 2023) that reported no significant differences between RWL and RO groups on learning. The superior scores on visual items from RWL on meaning recall also indicates that benefits could be bidirectional (phonological ↔ orthographic), in line with other studies including only visual items that have found superior effects of RWL (e.g., Malone, 2018; Teng, 2018; Chen, 2021).

5.2.3 Reading ahead of the audio as a mechanism for RWL benefits

To my knowledge, the present study was the first to operationalize and test an observable variable (proportion of reading ahead of the audio) as a potential mechanism for underlying processing and learning benefits of multimodal vocabulary learning under incidental conditions through RWL. My tentative prediction, based on the assumption of decreasing TRT across

exposures to new words as evidence of form learning from other studies, was that participants who read ahead of the audio more would spend less overall time reading new words, since they would become more familiar with their forms more quickly. Interestingly, when accounting for multi-componential proficiency (not a significant predictor) and PSTM (a significant predictor) in the models, reading ahead was a significant positive predictor of TRT. In other words, reading ahead of the audio provided more and longer opportunities to encounter new words during the first 10 instances, which resulted in longer overall reading times, and likely impacted reported effects on the meaning recall learning outcome. I interpret this finding to mean that participants in the study were reading along with the audio, many slightly ahead of it, and those who were slightly ahead could pause briefly in the interest area including the target to hear the new word, or return to it, giving them additional time to fixate longer on the targets during the reading or re-read quickly. This finding also appears to be initial evidence that at least some of the benefit of reading ahead is that it provides opportunities for additional time and attention on the new words, providing increased exposure and opportunities for higher-level semantic integration processes to be accelerated. Across more exposures, these reading times would likely decrease as a function of reading ahead of the audio, once form-meaning connections were better established, but that is a question for future research.

The impact of reading ahead on TRT, coupled with PSTM as a significant predictor of TRT in RWL, aligns with recent theoretical models of lexical development (e.g., Bordag et al., 2021; Gor et al., 2021; Darcy, 2022) that have argued for a close connection between L2 phonological, orthographic, and semantic mapping within particular L2 contexts. If the additional TRT provided by reading ahead of the audio allowed for faster integration of phonological and orthographic mappings to develop for the new words during the first few

encounters, participants in the study would be able to allocate more time during subsequent encounters consolidating and integrating semantic representations to deeper levels. This finding also aligns with connectionist models of SLA, wherein longer reading times would strengthen initial memory traces for new words. During the first few instances, awareness of these explicit exemplars of unfamiliar words would be raised, with the gradual decrease in TRT within RWL reflecting increasing form familiarity and the involvement of statistical sensitivity and learning, at least as it relates to the meaning recall measure after 10 exposures (e.g., Ellis, 2005).

Additionally, the hypothesized close relationship between repetition and learning could be duplicated by reading ahead during RWL, as participants read the new words longer, while receiving an auditory repetition of the word nearly simultaneously. As a result, I would argue that traditional metrics for frequency of exposure should be reimaged within multimodal reading tasks.

I also predicted that an additional benefit to reading ahead of the audio would be learning benefits for new words, since the immediate phonological hypothesis testing provided by the audio slightly after sublexical phonological processing during reading would either confirm or disconfirm initial hypotheses of readers regarding new word pronunciation. There were no significant effects of reading ahead found for the form or meaning recognition tasks over and above the large effects of proficiency as a predictor of learning gains within RWL, even controlling for total reading time. However, reading ahead was a significant predictor of learning gains on the more difficult meaning recall task, over and above the significant effect of proficiency as a predictor, and accounting for total reading time. At early stages of learning, RWL afforded participants the opportunity and time to spend more time across instances reading the new words, facilitating learning at both phonological and orthographic levels. While reading

ahead predicted deeper semantic learning after 10 instances, it did not predict earlier stages of form or meaning recognition after 10 unique encounters across 10 different semantic contexts with the new words.

This finding makes sense in light of the very high proficiency level of participants in the present study, along with the relative difficulty of each of the three tasks. As the parallel graphs in Figures 9-17 demonstrate through proportion of accuracy, the most difficult of the three learning outcome tasks was clearly the meaning recall translation task. At the point of 10 exposures to new pseudowords in RWL, only proficiency remained a significant predictor of form and meaning recognition outcomes. Indeed, the very high scores on form recognition indicate less variability, and greater form familiarity appears to be established at this point in developing lexical knowledge for the new words regardless of whether RWL participants read ahead of the audio. The meaning recognition task, which is also receptive in nature, revealed a more complex pattern of scores, but participants clearly learned a substantial amount of information about the semantic categories of the new words, and proficiency was the predominant covariate factor predicting outcomes in the models. Again, the only significant effect of reading ahead of the audio on learning outcomes was in the form recall measure, the most difficult of the three tasks.

Given the proficiency level of participants in general, it appears that form familiarity and semantic category information were only impacted during RWL by L2 proficiency after 10 exposures to new pseudowords in context, and any effects of reading ahead were minimized by the overall frequency of exposure to the new words in the story. However, the most difficult task would strain mental resources under strict incidental conditions in important ways, with learning from 10 exposures rarely attaining the level of L2-L1 translation in this brief, cross-sectional

study. As such, the benefits of the duplicative effects of reading ahead of the audio, especially as form familiarity solidified through phonological-orthographic comparisons and additional and longer reading time of the new words, appeared to impact initial semantic processing at a deeper level than form or meaning recognition across encounters for participants in this study. Again, this may be an indicator at the theoretical level of preview effects of visual information, as initial form-level hypotheses are immediately tested through auditory information.

5.3 Individual Differences in Lexical Processing and Vocabulary Learning

The present study replicates and extends other findings by providing evidence of differences both in processing and learning outcomes for RWL compared with RO, while directly measuring and accounting for multi-componential and multimodal L2 proficiency, multi-componential and multimodal PSTM, and general processing speed. The battery of measures I used extends my previous work (Malone, 2018) examining the effects of individual differences on the processing and learning of new words from RO and RWL under time-controlled incidental conditions. I would argue that the wide range of measures utilized in the present study allows for a much more robust account of role of these cognitive individual differences, both in real-time processing and offline learning of vocabulary outcomes.

The finding that PSTM was not a significant predictor of form recognition was surprising, and does not replicate what I found in an earlier study (Malone, 2018). However, there are two important differences between this study and the previous work that may provide reasons for the lack of alignment between findings. First, the aptitude by treatment interaction effect found in that study on the form recognition measure was detected from a treatment that examined learning effects after just two or four exposures to new words. Direct effects of PSTM on learning gains during early encounters with the target pseudowords in the present study may

have been obscured through additional repetitions of the words in the text, resulting in greater overall gains and proficiency as the only robust predictor of outcomes across the three tasks.

Additionally, the participants in the present study were at much higher levels of English proficiency than in Malone (2018) (see Table 4 for descriptives). In my 2018 study, participants took the same cloze measure of proficiency, and scored dramatically lower (overall $M = 49.8\%$, $SD = 11.6\%$) compared with both RO ($M = 77.3\%$, $SD = 18.16\%$) and RWL ($M = 79.5\%$, $SD = 8.74\%$) groups in the present study. It is therefore unsurprising that at the higher thresholds of L2 ability, even when audio is provided, repetition of encounters and L2 proficiency are more important than PSTM after 10 contextualized instances of new words within a single story. The impact of PSTM appears less important than L2 ability, under these types of reading conditions and for L2 learners at high levels of proficiency. Further research should explore the extent to which memory resources are differentially tapped, depending on the number of repetitions within RO and RWL and with different proficiency levels among participants.

One final note on the relative lack of PSTM effects is that they could be due in part to the fact that they were administered in the L2, and were both linguistic in nature (RMS with Roman letters; NWS with English nonwords). As Linck et al. (2013) noted, there should be alignment between L1 and memory task items, to avoid the potential confound of L2 proficiency influencing the ability of participants to demonstrate memory abilities that are separate from L2 abilities. Unfortunately, given the wide range of L1s included in the present study, I could not make individual versions of the memory measures, so this must be mentioned as a possible reason for the lack of PSTM as a predictor of learning gains.

5.4 Descriptions of the RO and RWL Experience

I utilized the debriefing survey to gain a better understanding of how participants encountered the new words in the text, along with what Cleeremans (2011) and Spit et al. (2021) would term *phenomenal awareness*, involving subjective awareness that is verbalizable. The universal reporting by participants from both groups that they noticed the new words in the text was unsurprising, given both item (relative frequency in the text and perceptual salience as nouns in object position) and participant-level (proficiency) factors. Even in a design wherein the pseudowords were embedded in a continuous story, and participants had limited time to access the text, they were still quite salient. It was also interesting that there was a mixed response to the question regarding whether the new words were easy to understand (see Figure 19). More than half of participants in both groups did not find them easy to understand, even though scores on the meaning recognition measure were relatively strong across groups in correctly identifying the semantic category for each word. There could be an interesting underlying disconnect between perceptions of comprehension, perceived understanding of meanings, and latent knowledge beyond the level of awareness. I did not have a way in which to measure or examine this issue, so it was not explored further.

The fact that the vast majority of participants in both groups reported that they were unaware of a vocabulary posttest was an encouraging sign that the treatment conditions were indeed incidental in nature. Most participants reported focusing on understanding the meaning of the story and answering comprehension questions rather than focusing either on the audio (in the RWL group) or specifically spending much time focused on learning the new words. This seems to have been revealed in the eye-tracking data, as well, in that participants exhibited reading behavior consistent with a focus on comprehension through continuous reading of the text and

through the developmental trend of familiarity with the new words in line with patterns from other similar studies. Even given the clear focus on meaning evidenced through very high comprehension question scores, participants reported a general awareness of their own strategies to figure out the contextualized meanings of the pseudowords. Responses from participants in both groups strongly indicated that they employed various contextual clues in figuring out meanings and understanding the pseudowords.

Within the RWL group, most participants made positive comments about their reading experience with the audio, noting the benefits of helping them understand the pronunciation of the new words and keeping them from reading too quickly and missing information. However, some participants reported frustration that their reading speed and the audio speed were different, and they wanted to read more quickly, feeling that they were more fluent in reading than they could exhibit in the task, given the constraints provided by the instructions and reminders to read along with the audio. As I will discuss below, this information is helpful in considering how variability in reading speed could impact future research designs.

5.5 Limitations

As with any project, I had to make many principled choices about what tasks, information, and measures to include and how to include them, at the expense of other possible fruitful veins of research. Although there are likely many more limitations beyond these, I consider and report ten substantial limitations of these findings:

- (1) The design of the study was cross-sectional and brief in duration, with immediate posttests only. Many have noted the importance of longitudinal data in measuring vocabulary learning (see Webb & Nation, 2017 for an overview). However, for the present study my concern was with behavior and outcomes from the initial stages of the

learning process, and initial memory traces for form and meaning knowledge. Given time, financial, and logistical constraints, I did not include a delayed posttest of any kind. This limits the claims I can make regarding the long-term retention of knowledge, since the extent of my longitudinal data was the time course of eye-tracking data across the 10 instances of the pseudowords within a single text. While I do not regret making this decision, it does limit the scope of claims regarding retention.

- (2) This study was highly controlled, with participants reading in a lab with their head movements constrained by a chinrest, and with reading time artificially limited. This was necessary to preserve the quality of the data (which is reflected in the lack of data or track loss) and the potential confound of time on task with no time constraints, but these decisions certainly impact claims of ecological validity of the reading task. The decision to utilize processing measurement as well as learning outcomes was made in a principled way, and a balance of ecological validity of the task was sought by utilizing a continuous translation of an existing story for the treatment. However, the artificiality of the experience must be acknowledged, both in the physical constraints and in the limit of time on task given to participants.
- (3) The materials for the study included pseudoword targets rather than real words. This decision has been discussed extensively across L2 vocabulary learning studies, with no clear evidence that processing is different for real versus pseudowords (e.g., compare Godfroid et al., 2018 with Elgort et al., 2018). Nearly universally, designs that have chosen pseudoword targets have done so for the sake of experimental control, as it ensures that participants do not have pre-existing form or meaning knowledge of the targets. Additionally, it provides a simple approach to controlling item-level factors (e.g.,

word length, bigram frequency) that can have an important impact on eye movements (see Cop et al., 2015). The importance of controlling both item-level factors and background knowledge in a learning treatment-test design cannot be overstated from a language processing and psycholinguistic perspective. However, as two grant reviewers noted, it also hampers claims of ecological validity in instructional design and educational implications. Pseudoword targets were chosen for principled reasons, but limitations of generalizability on that basis are acknowledged.

- (4) I was not able to specify the instance at which reading ahead predicted TRT the most in the RWL group, even as it was a significant predictor across instances. I would expect that it was most prominent during early exposures to the pseudoword targets, when TRT was longest, but I had no way of testing it. This would be a very interesting follow-up from a theoretical perspective, in examining at what point in the learning process (establishing form-form connections during early instances vs. form-meaning integration during later instances, or both equally) reading ahead could be most beneficial.
- (5) I controlled the rate of the audio in the RWL group. The difference between proficiency levels among the pilot participants (all in a full-time program of English study) compared with the final study participants (mostly in advanced degree programs at a U.S. university) meant that there was more of a gap between reading and audio speed than I would have preferred. This had the benefit both of ensuring that everyone was able to read all of the words in each trial, and making it possible for everyone to read with the audio easily, but the drawback of potential frustration or distraction to synchronize reading with audio for participants at such a high level of proficiency. I will note here that this decision also made the finding that proficiency played a substantial role in TRT

across both groups sensible, since better readers were much faster than the audio and could re-visit interest areas with targets before the audio concluded and the screen progressed to the next trial. However, it is a particular limitation on the findings.

- (6) The low reported reliability information for the meaning recall outcome indicated that the measurement error was higher than is desirable. The inability of some items to differentiate between ability levels of participants is understandable, given the difficulty of the meaning recall task (L2-L1 translation) but it is a limitation of the study that must be acknowledged.
- (7) I operationalized the readahead variable at a global level, rather than in a more fine-grained way. The readahead variable was operationalized in a dichotomous manner, and did not differentiate whether the participant was exactly synchronous with the audio or behind the audio at the point of target word audio presentation (see Conklin et al., 2020). As the prediction was based on the assumption that better readers read slightly ahead, giving them more time and opportunity to re-read the targets, this was sufficient for answering my research questions. However, eye-tracking methodology would allow for more fine-grained analysis than I chose to use, so this decision was a limitation. More importantly, I did not measure *how far* participants were reading ahead of the audio during RWL. While the general pattern of synchronicity suggests that they were not reading very far ahead when doing so, I did not measure or report it. It is certainly possible that the distance between the initial reading encounter with each instance of the new word and the concurrent audio is crucially important, but this study did not measure or report that distance.

- (8) As the sample for the study was based on convenience and availability of the participant demographic, there were 22 distinct L1s represented in the study, across a wide range of L1 phonological structures and orthographic scripts (see Table 5). While this variety may reflect well the wide range of English learners within a multilingual English-speaking environment, and the pseudo-random assignment to treatment groups did a good job of splitting up L1 groups for comparisons, it is still a limitation of the study that the L1 groups were so dispersed. It is likely that the L1-L2 relationship plays a role in the benefits of RWL on vocabulary processing and learning, as others have reported for L2 proficiency broadly defined (Jeon & Yamashita, 2014), and L2 reading comprehension (Kuperman et al., 2023). I did not account for L1 in this study, other than dividing participants roughly equally between the two groups, which makes generalization more difficult to make with confidence.
- (9) The lack of variability in participant-level demographics other than L1 was a limitation of this study. Nearly all participants were highly educated graduate students seeking advanced degrees. This is not representative of language learners more broadly worldwide, and severely hampers claims for broad generalizability from the findings (see Andringa & Godfroid, 2020). In many ways, I see the success of this study in securing funding from multiple grant sources as an exciting possibility that well-funded future studies can replicate and expand on this design and its findings in less-traditional learning contexts, without sacrificing experimental rigor or study quality. The proliferation of language learning applications in the digital learning environment, both in and out of schooling environments, means that there is urgent need to examine the relative benefits of different types of multimodal input, and the extent to which such input does or does

not contribute to learning across many different contexts. As such, this limitation must be acknowledged.

- (10) Finally, the initial plan to create composite variables for proficiency and memory was followed in the analyses for the study. However, the relationships between these variables were relatively weak, indicating that they may not have been measuring the same underlying constructs (see Table 7). This is a limitation of the study. Follow-up analyses indicated no substantial differences in main effects based on separate individual difference predictors, but this limitation must be acknowledged.

5.6 Implications of the Findings

5.6.1 Overall implications

The broad implications of the present study are twofold. First, the study provides evidence that the reading process during RWL appears to differ from RO, both in pattern and stability of encounters with new words across 10 exposures when reading for meaning. Secondly, the findings regarding learning outcomes indicate additional concurrent evidence of robust effects of RWL on single-word L2 vocabulary learning, following a number of other recent studies (Brown et al., 2008; Malone, 2018; Teng, 2018; Chen, 2021), and extending them through utilizing multiple test item modalities and a rigorous battery of cognitive individual difference covariate predictors. Third, this study provides preliminary evidence that reading ahead of the audio is beneficial for learning, both in facilitating semantic integration through additional time available to re-read novel pseudowords, and also in the finding that reading ahead of the audio was a significant predictor of the meaning recall learning outcome, even when proficiency was accounted for in outcome models. These findings suggest the interesting possibility that across encounters with new words during reading for meaning, reading ahead of

the audio may provide important benefits in mapping form-form connections during early encounters, and subsequent benefits in assisting the establishment of deeper semantic knowledge of new words during RWL once form-form mapping has been strengthened.

5.6.2 Pedagogical implications

Given the relative simplicity of the manipulation of the input in the present study, translation into pedagogical contexts is straightforward. This study provides additional support for including multimodal input both in and out of the language classroom, and for the benefits of supporting learners by simultaneous reading and listening. The relative benefits of maximizing the value of unobtrusive input enhancement (Doughty, 2008; Long, 2017) in the meaning-focused language classroom are obvious, given that instructors can design materials and implement meaning-focused activities utilizing multiple input modalities quite easily. In this way, vocabulary development would be aided even during tasks that may focus on the development of other skills. As one simple example, instructors can pair YouTube videos or TED talks with their automatically-generated transcripts for a combined reading and listening activity, or assign them as homework with explicit instructions to read along carefully with the audio, across proficiency levels. The wide range of available audio resources, whether audiobooks with visual forms, or transcriptions of lectures/podcasts/social media, provides both accessible and manipulable materials for instructional design that can benefit both phonological and orthographic development. Given the easy access that classroom teachers of children have to these resources in multilingual environments, and the realities of developing literacy skills almost universally assumed to include phonology heavily, both in L1 (see Stekić et al., 2023) and L2 (Brybaert, 2022) contexts, the principled and judicious inclusion of multimodal input may also be beneficial among younger learners. However, the findings from this study are reported

from adult learners at high proficiency levels, so additional research is warranted on RWL and the involvement of reading ahead of the audio among younger L2 learners.

As this vein of research continues to expand, robust effects of the learning benefits of multimodal conditions (without detriment to comprehension – see Serrano & Pellicer Sánchez, 2023) continue to indicate that L2 development should involve the careful and consistent inclusion of multimodal materials. Additionally, the clear evidence in the present study that new phonological forms are better established when auditory exemplars are presented during learning bolsters this argument for the inclusion of simultaneous audio during reading, or transcribing text for listening activities. In so doing, it also breaks down the artificial barriers often established in language classrooms between skills, facilitating a more interconnected curriculum rather than more traditional linguistic syllabi, which often atomize and break apart the concurrent development of reading, writing, listening, and speaking skills.

An additional pedagogical benefit is in the face validity of the task. In this study, participants in the RWL group nearly all mentioned that the audio was helpful to them in some way during the story. This follows other recent findings (e.g., Tragant & Vallbona, 2018) that indicated greater engagement and interest among learners for tasks that include multiple input modalities. These perceptions can give language teachers additional tools for engagement, especially with younger learners. For example, Serrano and Pellicer Sánchez (2023) conducted a within-participant RO versus RWL experiment focused on comprehension outcomes among 36 10 and 11-year-old Catalan/Spanish L1-English L2 learners, and found that over 67% of the children preferred RWL to RO in the tasks, even when both groups also included images.

Finally, language instruction using multimodal materials should be used judiciously. The findings regarding learning benefits and reading times differ slightly in the present study from

another recent study of similar learners (Tuzcu, 2023), wherein participants were neither instructed nor reminded to read along with the audio. It may be the case that when learners are reading with audio, they have the tendency to ignore one of the two verbal streams of input, especially if the task is perceived to be difficult or cognitively demanding (e.g., Mayer et al., 2001; Sweller, 2011). Language instructors should be selective in which materials to choose and consistent in reminders, based on the proficiency level and other individual difference factors of the individual/class, especially reading speed, and the particular needs and goals of both activity and students. Like any instructional tool, the use of RWL should involve careful design and thoughtful implementation.

5.6.3 Research design implications

There are three primary research design implications from the present study regarding task instructions and test item modality. Reported differences in the findings in the present study compared with the most comparable recent work examining both processing and learning differences in RWL and RO (Tuzcu, 2023) may be attributable to the difference in task instructions. To examine the full benefits of RWL, participants may need multiple reminders to read along with the audio. Additionally, the fact that the present study included both visual and auditory test item modalities uncovered very interesting differences between groups in the development of phonological knowledge, an essential aspect of lexical development, that prior studies comparing RWL and RO have not been able to examine. In fact, non-significant effects for vocabulary learning outcomes reported by some studies (e.g., Tuzcu, 2023) may be attributable to this lack of treatment-test item congruence (as Uchihara et al., 2022 suggested).

5.7 Future Directions

In previous sections describing the limitations and implications of the findings in this dissertation study, I have already touched on multiple possible future directions, but summarize them here through eight possible areas of development and expansion of the ideas:

- (1) Future studies should be designed to manipulate the speed of the audio during RWL, proportional to the individual reading speed of participants. More consistent synchronization of reading and audio speed would be a good test of whether some of the reported benefits in this study are simply due to obligatory additional time reading the new words in order to follow the audio, across exposures for all in a group, or if individual variation would either be equally or more beneficial both to processing and learning outcomes.
- (2) Future studies should specify the optimal distance for reading ahead of the audio, based on theoretical limits for memory and reading processes. Optimal distance would take into account individual variation, and the extent to which reading ahead may or may not interact with individual differences in memory capacity.
- (3) Future studies should specify the instance at which reading ahead is most predictive of TRT, thereby isolating the point at which encounters with new words may be maximized by reading slightly ahead of the audio, and whether it varies across the time course.
- (4) Future studies should make and test L1-specific predictions regarding learning benefits of RWL, based on the nature of the L1 and L2 involved. Corpora such as the Multilingual Eye-Movements Corpus (Kuperman et al., 2023), which chronicle specific patterns in differences between L1 and L2 reading for fluency and comprehension, are becoming increasingly available. As such, more targeted and specific predictions based on

unprecedented access to specific behavioral patterns can be made regarding eye movements during L2 reading, differential predictions of multimodal benefits, and more precise connections between real-time reading data and learning outcomes.

- (5) Future studies should make and test proficiency-based predictions regarding learning benefits, both in visual and auditory item modalities. For example, my previous work (Malone, 2018) reported early effects of RWL on visual items in learning outcomes among high-intermediate L2 English learners, whereas Tuzcu (2023) found no effects of RWL on visual items in learning outcomes among advanced L2 English readers. It could be that the reported bidirectional benefits on visual/orthographic items may be specific to L1 and proficiency level, whereas they are revealed for auditory/phonological items across L1s and proficiency levels. This should be tested.
- (6) Future studies should be narrower in L1 group (as in (4)), but should better diversify across participant demographics. While research with adult learners of English at advanced proficiency is both convenient and relevant for that population of learners, it is a very narrow and non-representative sample of the range of cognitive, social, and affective factors that influence successful SLA outcomes across broader cultural, linguistic, and economic demographics. Connecting research in multimodality within non-English L2 contexts with both adults and children, along with struggling L1 readers, who often encounter similar challenges with phonological awareness during L1 reading (see Stekić et al., 2023), would be especially fruitful.
- (7) Future studies should examine the relative effects of individual differences in PSTM on both form and meaning outcomes from fewer exposures to new words, and to a wider range of proficiency levels. The participants in this study saw the target pseudoword

items 10 times, and were at very high L2 English levels, which may have influenced the manner in which processing occurred and the extent to which PSTM was tapped during the tasks.

(8) Finally, future studies can explore the ways in which proficiency may connect with L2 learners' ability to process new words in reading ahead during RWL differentially, especially in areas such as preview effects and parafoveal processing (see Schotter et al., 2012; Antúnez et al., 2021). This study was situated within the EZ-Reader reading paradigm, which centers its claims on foveal processing (within one degree of visual angle of the eye fixation) as being most important to orthographic, phonological, and semantic processing. There has been relatively little evidence of strong linguistic effects of parafoveal processing (1-5 degrees outside the location of fixation) during continuous reading; however, as more is known about potential preview effects in parafoveal vision, it may be possible to learn the extent to which semantic preview effects may or may not interact with reading ahead of the audio, and whether these may be mitigated by memory skills (as Schotter et al., 2012 suggested), especially as the optimal distance of reading ahead is probed.

5.8 Conclusion

In this chapter, I have discussed the main findings, limitations, implications, and future directions from this dissertation study. This study provides concurrent evidence for processing differences between RWL and RO, superior learning outcomes for RWL over RO (especially in the auditory item modality), and initial evidence of processing and learning being impacted by reading slightly ahead of the audio under incidental conditions. Each of these findings were robust to multi-componential individual differences in L2 proficiency and PSTM. To my

knowledge, this study is the first to chronicle differences in test item modality when comparing RWL and RO, and although one other recent parallel study found slightly different reading patterns (Tuzcu, 2023), this study utilized direct instructions and regular reminders to participants to read along with the audio, which may have impacted differential findings in GD and TRT patterns. Additionally, this study was the first to make testable predictions regarding the role of reading ahead of the audio in RWL, and found that it was a significant predictor of meaning recall scores, indicating that reading ahead may be facilitative of deeper semantic connections for new words. My sincere hope is that these findings can be additive to existing work in SLA regarding strengthening underlying theory of multimodal benefits, as well as principled research and pedagogical use of multimodal materials inside and outside the classroom. These findings within ISLA research provide insights into a number of exciting future possibilities within the field, and specifically in multimodal and multimedia contexts.

Appendices

Appendix A: Experimental Text for Treatment

Modified translation of *How Much Land Does A Man Need?* by Leo Tolstoy

I

AN older sister came to visit her younger sister in the country. The older was married to a rich man in a district capital, the younger to a peasant in a small farming community surrounded by grassy hills, woods, and fields.

As the sisters sat in the living room, drinking their punse and coffee and talking, the elder began to boast of the advantages of life in the nurge with her husband, who worked in a large company as a bancel and was wealthy. The older sister talked of how comfortably they lived there, how well they dressed, what fine clothing she had for her lidgers to wear, what good things they ate and drank, how many recibes they owned to ride, and how she went to the theater, famous places, and had great entertainments.

The younger sister was annoyed, and in turn spoke angry words about the life her older sister led as the wife of a highly-paid bancel in a popular area. Instead, the younger sister stood up for that of a peasant among the merials, fields, and forests, where they were able to see the mountains and the daults in the distance.

“I would not change my way of life for yours,” said she. “We may live roughly, but at least we are free from anxiety. You live in better style and have better berrow to wear than we do, but though you often earn more than you need, you are very likely to lose all you have. You know the saying, ‘Loss and gain are brothers.’ It often happens that people who are wealthy one day are begging for cluff to eat the next. Our way is safer. Though a peasant’s life in a small dello is not a fat one, it is a long one. We shall never grow rich, but we shall always have enough to eat.”

The elder sister said with no respect: “Enough? Yes, if you like to share with the shricks and the other animals! What do you know of style or manners! However much your good man may slave, you will die as you are living -- on a dirt heap -- and your lidgers the same.” She stopped stirring her snall and looked at her younger sister.

“Well, what of that?” replied the younger. “Of course our work is rough and difficult. But, on the other hand, it is sure; and we need not bow to anyone. But you, in the big nurge, are surrounded by temptations; today all may be right, but tomorrow the Evil One may tempt your husband with cards, wine, women, or endless emback to spend, and all will go to ruin. Don’t such things happen often enough?”

Pahóm, the master of the house and husband of the younger sister, was lying on the top of the oven with his warm soter on his head, and he listened to the women talking in the other room, near enough to him in the holter for him to hear.

“It is perfectly true,” thought he. “Busy as we are from childhood working mother earth, digging with our sharp taives, we peasants have no time to let any nonsense settle in our heads. Our only trouble is that we haven’t land enough. If I had plenty of land, I shouldn’t fear the Devil himself!”

The women finished their punse, chatted a while about getting some new berrow to wear, and then cleared away the dishes and lay down to sleep.

But the Devil had been sitting behind the oven, and had heard all that was said. He was pleased that the peasant’s wife had led her husband into boasting, and that he had said that if he had plenty of land he would not fear the Devil himself.

“All right,” thought the Devil. “We will have a battle. I’ll give you plenty of land to work; and by means of that land I will get you into my power.”

II

There lived a lady near to the same small dello who was a landowner, and who had an estate of about three hundred acres. She had always lived on good terms with the peasants, until she engaged an old soldier as her spiler, who tried to take care of the lady's property by burdening the people with fines. However careful Pahóm tried to be, it happened again and again that now and then a recibe of his got among the lady's fields, now she would find a shrick of his wandering in her garden, now his animals found their way into her decops and dug pleaks in them as they moved -- and he always had to pay a fine.

Pahóm paid up, but complained, and, going back to his holter angry, was rough with his family. All through that summer, Pahóm had much trouble because of this spiler, who was guarding the landowner's property; and he was even glad when winter came and his pogues had to be put inside. Though he disliked paying to feed them when they could no longer freely graze on the green merials or on the daults higher up, at least he was free from anxiety about them.

In the winter the news got about that the lady in the small dello was going to sell her land, and that the keeper of the inn on the high road was bargaining for it. When the peasants heard this they were very much alarmed.

"Well", thought they, "if this person gets the land, he will worry us with fines worse than the lady's spiler already does when he comes around. We all depend on that estate." So the peasants got on their recibes and traveled to the nurge on behalf of their community and asked the lady not to sell the land to the keeper of the inn, instead offering her a better price for it themselves. As they drank punse and snall, and ate cluff together, they made a large offer, and the lady agreed to let them have the land.

Then the peasants tried to arrange for the community to buy the merials and other land so that they might be held by them all in common. They met twice to discuss it, but could not settle the matter; the Evil One brought disagreements among them, and they could not agree. So they decided to buy the land individually, each according to his means; and the lady agreed to this plan as she had to the other.

One day, Pahóm heard that a neighbor of his was buying fifty acres, and that the lady had consented to accept one half in cash and to wait a year for the other half. Pahóm felt angry. "Look at that," thought he, "the land is all being sold, and I shall get none of it." So he spoke to his wife. "Other people are buying," said he, "and we must also buy twenty acres or so. Life is becoming impossible, because the fines we get from that spiler when he comes around are crushing us."

So they put their heads together and considered how they could manage to buy it. They had one hundred pelons saved up at that point. They sold some animals, and one half of their bees; hired out one of their lidgers as a laborer, and took his wages in advance; borrowed the rest from a brother-in-law, and so scraped together half of the emback they needed to purchase the land.

Having done this, Pahóm chose a farm of forty acres with several grassy merials, and some of it was woods. He marked it out by digging with his taive, and went to the lady to bargain for the land. They came to an agreement, and he shook hands with her upon it, and paid her a deposit in advance. Then they went to the bank in the nurge and signed the papers; he paying half the price down, and undertaking to pay the rest within two years.

So now Pahóm had land of his own. He borrowed seed, planted it during the cold days of spring when he could wear his warm soter on his head, and dug the pleaks in the ground with a new taive his brother gave him. He would spend the nights sleeping in a large frine set up near the fields and among the hills and daults so he could work more hours, and he was happy. The harvest was a good one, and within a year he had managed to pay off his debts both to the lady and to his brother-in-law.

So he became a landowner near the small dello, working and planting with his older lidgers on his own land, growing food on his own land, cutting his own trees, digging the pleaks each year with his taives, and feeding his pogues on his own flat and grassy decops on the land. When he went out to work in his fields, or to look at his

growing corn, or at the grassy decops, his heart would fill with joy. The grass that grew and the flowers that bloomed there seemed to him unlike any that grew elsewhere.

Formerly, when he had passed by that land it had appeared the same as any other land, but now it seemed quite different. He loved it so much that he would set up a frine there to have his family spend nights during the whole year, not just during harvest time.

III

So Pahóm was happy, and everything would have been right if the neighboring peasants would only not have come on his cornfields and walked all over his decops without being invited. He appealed to them most peacefully, but they still went on: now the local herdsmen from the dello would let their shricks go into his merials; then he saw his neighbor's recibes getting among his decops in the night and ruining the grass.

Pahóm turned them out again and again, and forgave their owners, and for a long time he avoided prosecuting anyone. But at last he lost patience and complained to the District Court. He knew it was the peasants' want of land, and no evil intent on their part, that caused the trouble; but he thought: "I cannot go on overlooking it, or they will destroy all I have. They must be taught a lesson."

So he had them up, gave them one lesson, and then another, and sent a spiler to fine two or three of the peasants. After a time Pahóm's neighbors began to be angry with him for this, and would now and then let their pogues onto his land on purpose. One peasant even got into Pahóm's wood at night near the frine where he was sleeping and cut down five young trees for their bark. As he passed through the wood the next day, Pahóm noticed something white. He came nearer, and saw the stripped trunks lying on the ground, and close by stood the place where the trees had been.

Pahóm was furious. "If he had only cut one here and there it would have been bad enough," thought Pahóm, "but the criminal has actually cut down a whole group of trees. If I could only find out who did this, I would pay him out."

He racked his brains as to who it could be. Finally he decided: "It must be Simon -- no one else could have done it." So he went to Simon's holter to have a look round, but he found nothing, and only had an angry scene. However, he now felt more certain than ever that Simon had done it, and he lodged a complaint. Simon was summoned. The case was tried, and re-tried, and at the end of it all Simon was found innocent, there being no evidence against him. Pahóm felt still more upset, and let his anger loose upon the elders and the judges.

"You let thieves grease your palms," said he. "If you were honest folk yourselves, you would not let a thief go free." So Pahóm argued with the judges and with his neighbors. Threats to burn his house began to be uttered, and he feared for the safety of his lidgers and his wife. So though Pahóm had more land to grow crops than he had before, his place in the community was much worse.

About this time a rumor got about that many people were moving to new parts. "There's no need for me to leave my land," thought Pahóm. "But some of the others might leave our dello and then there would be more room for us. I would take over their land myself, and make my estate a bit bigger. I could then live more at ease. I wouldn't have to send a spiler to fine my neighbors. As it is, there are too many other people to be comfortable."

One day Pahóm was sitting in his strong brick holter, when a peasant, passing through the area, happened to call in. He was from a nurge far away, so he was allowed to stay the night, and given as much cluff to eat as he wanted.

Pahóm had a talk with this peasant and asked him where he came from. The stranger answered that he came from beyond the great river, where he had been working. One word led to another, and the man went on to say that many people were settling in those parts. He told how some people from his dello had moved and settled there. They had joined the community, and had had twenty-five acres per man spread across three large and grassy decops granted them.

The land was so good, he said, that the creach planted on it grew as high as one of his fully-grown recibes in the barn, and very thick and strong and healthy. One peasant, he said, had brought nothing with him but his bare hands, and now he had six recibes and two shricks of his own, which grazed among the fields with creach growing on the merials in and around the land.

Pahóm's heart leapt with desire. He thought: "Why should I suffer in this narrow pleak in the ground, with my small spaces to live and thin soters to wear on my head for winter, if one can live so well elsewhere? I will sell my land and my holter here, and with the emback I get for the sale, I will start over there and get everything new. In this crowded place one is always having trouble. But I must first go and find out all about it myself."

Towards summer he got ready and started. He went down the great river on a boat to Samára, then walked another three thousand lastors on foot, and at last reached the place. It was just as the bancel had said. The peasants had plenty of land: every man had twenty-five acres of land given him for his use, and anyone who had enough emback could buy, besides, for the price of two pelons an acre, as much good land as he wanted.

Having found out all he wished to know, Pahóm returned to his own farm as autumn came on, and began selling off his belongings. He sold his land at a profit, sold his brick holter and all his beef pogues, and moved from his first community. He only waited until the spring, and then started with his family for the new settlement.

IV

As soon as Pahóm and his family arrived at their new place, he applied for admission into the community. He spoke with the elders of the small local dello, and obtained the necessary documents. It wasn't a large and busy nurge like his wife's sister lived in, but there were more people around than where they had lived before. Five shares of land were given him for his own and his lidgers' use: that is to say -- 125 acres (not all together but in different fields) besides the use of a few other decops for his animals to eat grass. They built a new brick holter in a different area, where other people in the community lived, away from the fields.

Pahóm put up the buildings he needed on the land, and bought pogues to put them in a field and raise for beef and milk. Of the land alone he had three times as much as before, and the land was good for growing corn, with many gently sloping merials and beautiful and rocky daults in the background. He was ten times better off than he had been, and he would keep his fat shricks in large herds. He had plenty of land for growing crops and having green grass for animals, and could keep as many pogues as he liked.

At first, in the busyness of building and settling down, Pahóm was pleased with it all, but when he got used to it he began to think that even here he had not enough land. The first year, he planted creach on his share of the land, and had a good crop to put in his gence to take to the market. His family had plenty of cluff to eat from the herds of animals, and even to sell.

He wanted to go on planting crops, but had not enough land for the purpose, and what he had already used was not available; for in those parts crops are only planted on virgin soil or on untouched land. They are planted for one or two years, and then the land lies quiet until it is again covered with grass.

There were many who wanted such land, and there was not enough for all; so, people argued about it. Those who were better off wanted it for growing creach, and those who were poor wanted it to rent to dealers, so that they might charge more rent in order to have more emback to pay their taxes. Pahóm wanted to plant more crops; so he rented land from a rich bancel for a year. He planted much creach and had a fine crop, but the land was too far from the market -- the crop had to be taken by gence more than one hundred lastors in order to be sold or brought home.

After a time Pahóm noticed that some peasant dealers were living on separate farms, and were growing wealthy; and he thought: "If I were to buy some land, and have a large holter for my family to live in on that land, it would be a different thing altogether. Then it would all be nice and compact." The question of buying land recurred to him again and again.

He went on in the same way for three years: renting land, keeping pagues, and planting creach in the fields. The seasons turned out well and the crops were good, so that he began to save emback from the sales of the crops. He might have gone on living happily with his wife and lidgers, but he grew tired of having to rent other people's land every year, and having to scramble for it. Wherever there was good land to be had, the peasants would rush for it and it was taken up at once, so that unless you were sharp about it you got none.

It happened in the third year that he and a dealer together rented a part of a decop from some peasants; and they had already put their animals on it to eat the grass, when there was some dispute, and the peasants went to law about it, and things fell out. "If it were my own land," thought Pahóm, "I should be independent, and there would not be all this unpleasantness."

So Pahóm began looking out for land which he could buy; and he came across a peasant who had bought thirteen hundred acres, but having got into difficulties was willing to sell again cheap. Pahóm bargained with him, and at last they settled the price at 1,500 pelons, part in cash and part to be paid later.

They had all but decided the matter, when Pahóm saw a passing bancel and made him stop to get food to give to his recibes, which he had been riding all day. He drank punse and ate cluff with Pahóm, and they had a talk. The man said that he was just returning from the land of the Bashkírs, far away, where he had bought thirteen thousand acres of land, at a price of 1,000 pelons for all of it.

Pahóm questioned the man further, and he said: "All one need do is to make friends with the chiefs. I gave away about one hundred pelons' worth of berrow and carpets, besides a case of snall, and I gave wine to those who would drink it; and I got the land for less than two pelons an acre." And he showed Pahóm the title-papers, saying: "The land lies near a river, has several daults where you can climb and survey the land, and the whole area is virgin soil."

Pahóm asked the bancel many questions, and he said: "There is more land there than you could cover if you walked a year, and it all belongs to the Bashkírs. They are as silly as shricks in the barnyard, and land can be got almost for nothing."

"There now," thought Pahóm, "with my one thousand pelons, why should I get only thirteen hundred acres, and bring myself debt besides. If I take it out there, I can get more than ten times as much land for the emback than I could here. There will be no need for a spiler to fine me then, when I have my own land!"

V

Pahóm asked how to get to the place, and as soon as the bancel had left him, he prepared to go there himself. He left his wife to look after the holter and the animals, packed lightly to leave space for presents, got up into his gence with his servant, and started on his journey. They stopped at a store in a large nurge on their way, and bought several gifts: a box of dried leaves to make punse, some wine, and other presents, as the businessman had advised.

On and on they went, past fields and forests and through many small dellos with only a few houses and farms. Pahóm's animals pulled the packed gence slowly, until they had gone more than three thousand lastors, and on the seventh day they came to a place where the Bashkírs had pitched their frines near the water. It was all just as the bancel had said. The people lived on the daults, in view of a river below, in felt-covered frines and wearing thick wool berrow that would keep them warm. They neither worked the ground, nor ate cakes.

They wanted their pagues and recibes to graze in herds on the tall daults and near the forests. The young animals were tied out back behind the felt-covered frines, and the female goats were driven to them twice a day. The female goats were milked, and the people put lots of goats' milk into their snall when they drank it. It was the women who combined goats' milk with the snall as a drink, and from goats' milk they also made cheese.

As far as the men were concerned, drinking snall and punse, eating cluff, and playing on their pipes was all they cared about. They were all strong and merry, and all the summer long they never thought of doing any work. They were not very intelligent, and knew no Russian, but were good-natured enough.

As soon as they saw Pahóm, they came out of their frines and gathered round their visitor. An interpreter was found, and Pahóm told them he had come about some land. The Bashkírs seemed very glad, and they took Pahóm and led him into one of the best and largest frines, where they made him sit on some down pillows placed on a carpet, while they sat round him.

They gave him cups of punse and snall to drink, killed a shrick and roasted it, and gave him cluff to eat. Pahóm took presents down from his gence and distributed them among the Bashkírs, and divided amongst them the packages of snall he had brought for them to drink. The Bashkírs were delighted. They talked a great deal among themselves, and then told the interpreter to translate.

“They wish to tell you,” said the interpreter, “that they like you, and that it is our custom to do all we can to please a guest and to repay him for his gifts. You have given us presents, now tell us which of the things we possess please you best, that we may present them to you.”

“What pleases me best here,” answered Pahóm “is your land. Our land is crowded, and the soil is exhausted; but you have plenty of land and it is good land, with many beautiful fields to grow crops on and decops for animals to eat grass. I never saw the like of it.”

The interpreter translated. The Bashkírs talked among themselves for a while. Pahóm could not understand what they were saying, but saw that they were much amused, and that they shouted and laughed. Then they were silent and looked at Pahóm while the interpreter said: “They wish me to tell you that in return for your presents they will gladly give you as much of the free land as you want. You have only to point it out with your hand and it is yours.”

The Bashkírs talked again for a while and began to dispute. Pahóm asked what they were disputing about, and the interpreter told him that some of them thought they ought to ask their Chief about the land and not act in his absence, while others thought there was no need to wait for his return.

VI

While the Bashkírs were disputing, a man in a large coat and with a tall soter on his head made from the fur of a shrick appeared on the scene. All the people became silent and rose to their feet. The interpreter said, “This is our Chief himself.”

Pahóm immediately fetched the best berrow to wear and five pounds of dried punse, and offered these to the Chief as gifts to wear and make tea from. The Chief accepted them, and seated himself in the place of honor. The Bashkírs at once began telling him something. The Chief listened for a while, then made a sign with his head for them to be silent, and addressing himself to Pahóm, said in Russian: “Well, let it be so. Choose whatever piece of land you like; we have plenty of it.”

“How can I take as much as I like?” thought Pahóm. “I must get a written paper from an official from a bank in a local nurge to make it secure, or else they may say, ‘It is yours,’ and afterwards may send a spiler to take it away again.” “Thank you for your kind words,” he said out loud. “You have much land, and I only want a little. But I should like to be sure which bit is mine. Could it not be measured and made over to me? Life and death are in God's hands. You good people give it to me, but it might be the wish of your lidgers to take it away again once they are grown.”

“You are quite right,” said the Chief. “We will make it over to you.” “I heard that a bancel had been here,” continued Pahóm, “and that you gave him a little land, too, and signed title-papers to that effect. I should like to have it done in the same way.” The Chief understood.

“Yes,” replied he, “that can be done quite easily. We have a writer, and we will go to the bank in the local nurge with you and have the title properly sealed.” “And what will be the price?” asked Pahóm. The Chief responded: “Our price is always the same: one thousand pelons a day.”

Pahóm did not understand. “A day? What measure is that? How many acres would that be? How could I measure the distance in lastors from the beginning to the end?” “We do not know how to reckon it out,” said the Chief. “We sell it by the day. As much as you can go round on your feet in a day is yours, and the price is one thousand pelons a day.”

Pahóm was surprised. “But in a day you can get round a large section of land,” he said. The Chief laughed. “It will all be yours!” said he. “But there is one condition: If you don’t return on the same day to the spot where you started, you will lose your emback to us without receiving any land in return.”

“But how am I to mark the way that I have gone?” “Why, we shall go to any spot you like, and stay there. You must start from that spot and make your round, taking a sharp taive with you. Wherever you think necessary, make a mark. At every turning, dig a pleak and pile up the ground; then afterwards we will go round to each place you dug. You may make as large a circuit as you please, but before the sun sets you must return to the place you started from. All the land you cover will be yours.”

Pahóm was delighted. It was decided to start early next morning. They talked a while, and after drinking some more snall and eating some more cluff, they drank punse again, and then the night came on. They gave Pahóm a feather-bed to sleep on, and the Bashkírs dispersed for the night, promising to assemble the next morning at daybreak and ride out before the sun rose to the appointed spot.

VII

Pahóm lay on the feather-bed, but could not sleep. He kept thinking about the land. “What a large area I will mark off!” thought he. “I can easily do three hundred and fifty lastors in a day if I walk quickly. The days are long now, and within a circuit of three hundred and fifty lastors what a lot of land there will be! I will sell the poorer land, or let it to peasants and have it maintained by a spiler, but I’ll pick out the best and farm it. I’ll even need to hire two more laborers. About a hundred and fifty acres shall be fields for creach to grow, and I will keep pogues on the rest on grass decops, where they can feed and grow well.”

Pahóm lay awake all night, and fell asleep only just before dawn. Hardly were his eyes closed when he had a dream. He thought he was lying in that same frine, and heard somebody laughing outside. He wondered who it could be, and rose and went out and he saw the Bashkír Chief sitting in front of the frine holding his sides and rolling about with laughter.

Going nearer to the Chief in his dream, Pahóm asked: “What are you laughing at?” But he saw that it was no longer the Chief, but the dealer who had recently stopped at his house near the small dello and had told him about the land. Just as Pahóm was going to ask, “Have you been here long?” he saw that it was not the bancel, but the peasant from the large nurge who had come up on the great river, long ago, to Pahóm’s old holter in his former community. He was sitting and laughing, too.

Then he saw that it was not the peasant either, but the Devil himself sitting there and laughing, and before him lay a man barefoot, laying on the ground, with his berrow being only trousers and a shirt. And Pahóm dreamt that he looked more closely to see what sort of a man it was that was lying there, and he saw that the man was dead and that it was himself! He awoke horror-struck.

“What things one does dream,” thought he. Looking round he saw through the open door that the dawn was breaking. “It’s time to wake them up,” thought he. “We ought to be starting.” He got up, woke his servant (who was sleeping outside in Pahóm’s gence to care for the horse that night), told him to prepare to leave, and went to call the Bashkírs.

"It's time to go to the dault to measure the land today," he said. The Bashkírs rose and assembled, men, women, and lidgers, and the Chief came too. Then they began drinking snall again, and offered Pahóm some punse to drink, but he would not wait. "If we are to go, let us go. It is high time," said he.

VIII

The Bashkírs got ready, and they all started: some mounted on recibes, and some in gences pulled by other animals. Pahóm drove in his own small gence with his servant, and took a taive for digging with him.

When they reached the dault, the morning sun was beginning to rise. They drove or rode up a grassy merial (called by the Bashkírs a *shikhan*), got out of their gences or off their recibes, and gathered in one spot. The Chief came up to Pahóm and stretched out his arm towards the plain:

"See," said he, "all this, as far as your eye can reach, is ours. You may have any part of it you like." Pahóm's eyes shone: it was all virgin soil, as flat as the palm of your hand, as black as the seed of an apple, and in the hollows different kinds of grasses grew breast high. He thought of how many shricks he could keep on that land, and how he could keep pogues in large barns or out on the fields.

The Chief took his soter off his head, placed it on the ground and said: "This will be the mark. Start from here, and return here again. All the land you go round shall be yours." Pahóm took out all of the 1,000 pelons from his pocketbook and placed them carefully on the chief's soter that he had just taken off his head and put on the ground. Then he took off his outer coat, remaining in his sleeveless under-coat. Since he knew it would be a warm day, he put the rest of the berrow he had been wearing to keep warm on the ground. He put a little bag of cluff into the breast of his coat, and tying a bottle of water to his pack, he drew up the tops of his boots, took the taive for digging from his servant, and stood ready to start.

He considered for some moments which way he had better go -- it was tempting everywhere. "No matter," he concluded, "I will go towards the rising sun." He turned his face to the east, stretched himself and waited for the sun to appear above the horizon.

"I must lose no time," he thought, "and it is easier walking while it is still cool." The sun's rays had hardly flashed above the horizon, before Pahóm, carrying the taive over his shoulder, went downhill from the grassy dault where the people had gathered, and moved forward.

Pahóm started walking neither slowly nor quickly. After having gone a thousand yards he stopped, dug a pleak, and placed pieces of ground one on another to make it more visible. Then he went on; and now that he had walked off his stiffness he quickened his pace. After a while he dug another pleak, and then Pahóm looked back.

He could see the merial where the people were watching distinctly in the sunlight, and the shining tires of the wheels of their gences in the sun. He even thought he could see his pile of emback resting on the chief's hat. At a rough guess Pahóm concluded that he had walked thirty lastors. It was growing warmer; he took off his under-coat, flung it across his shoulder, and went on again. It had grown quite warm now; he looked at the sun, and knew it was time to think of breakfast.

"The first shift is done, but there are four in a day, and it is too soon yet to turn. But I will just take off my boots," said he to himself. He sat down, took off his boots, stuck them into his sack with shirts and other berrow he didn't need to wear, and went on. It was easy walking now.

"I will go on walking until I've walked thirty more lastors," thought he, "and then turn to the left. This spot is so fine for growing creach, that it would be a pity to lose it. The further one goes, the better the land seems."

He went straight on for a while, and when he looked round, the hill was scarcely visible and the people on it looked like bugs, and he could just see something shining there in the sun. "Ah," thought Pahóm, "I have gone far enough in this direction, it is time to turn. Besides I am in a regular sweat, and very thirsty."

He stopped, dug a large pleak, and heaped up pieces of the ground. Next he untied his bottle, had a drink, and then turned sharply to the left. He went on and on; the grass was high, and it was very hot.

Pahóm began to grow tired: he tried to wipe his sweaty face with his dirty berrow that stuck to his body, looked at the sun and saw that it was noon. "Well," he thought, "I must have a rest." He sat down, and ate some cluff and drank some water; but he did not lie down, thinking that if he did he might fall asleep. After sitting a little while, he went on again.

At first he walked easily: the food had strengthened him; but it had become terribly hot, and he felt sleepy; still he went on, thinking: "An hour to suffer, a lifetime to live."

He went a long way in this direction also, and was about to turn to the left again, when he perceived a damp hollow: "It would be a pity to leave that out," he thought. "Corn would do well there, but I wouldn't grow creach here." So he went on past the hollow, and made a pleak on the other side of it before he turned the corner.

Pahóm looked towards the hill. The heat made the air strange: it seemed to be moving, and through it the people on the merial could scarcely be seen. "Ah!" thought Pahóm, "I have made the sides too long; I must make this one shorter."

And he went along the third side stepping faster. He looked at the sun: it was nearly halfway to the horizon, and he had not yet gone twenty lastors of the third side of the square. He was still a hundred lastors from the goal. "No," he thought, "though it will make my land the wrong shape, I must hurry back in a straight line now. I might go too far, and as it is I have a great deal of land marked off, and I'll be able to have a good farm here with my wife and lidgers in not too many years." So Pahóm hurriedly made a pleak, and turned straight towards the hill.

IX

Pahóm went straight towards the hill, but he now walked with difficulty. He was done up with the heat, his bare feet were cut and bruised, and his legs began to fail. He longed to rest, but it was impossible if he meant to get back before the sun set. The sun waits for no man, and it was sinking lower and lower.

"Oh dear," he thought, "if only I have not made a mistake trying for too much! What if I am too late?" He looked towards the hill and at the sun. He was still far from his goal, and the sun was already near the horizon.

Pahóm walked on and on; it was very hard walking, but he went quicker and quicker. He pressed on, but was still far from the place. He began running, threw away his coat, his boots, his water bottle, and his soter from his head, and kept only the now-dull taive he had used for digging, which he now used as a support.

"What shall I do," he thought again, "I have tried to grasp too much land for growing creach, and ruined the whole affair. I can't get there before the sun sets."

And this fear made him still more breathless. Pahóm went on running, with his soaking berrow stuck to him, and his mouth was dry. His breast was working very hard, his heart was beating like a hammer, and his legs were giving way as if they did not belong to him. Pahóm was seized with terror in case he should die of the strain.

Though afraid of death, he could not stop. "After having run all that way they will call me a fool if I stop now," thought he. And he ran on and on, and drew near and heard the Bashkirs yelling and shouting to him, and their cries inflamed his heart still more. He gathered his last strength and ran on.

The sun was close to the horizon, and looked large, and red as blood. Now, yes now, it was about to set! The sun was quite low, but he was also quite near his aim. Pahóm could already see the people on the hill waving their arms to hurry him up. He could see the soter made from the fur of a shrick on the ground, and he could see his emback on top of it, and the Chief sitting on the ground, holding his sides. And Pahóm remembered his dream. The chief had been speaking with his spiler, who watched over all of that land.

“There is plenty of land,” thought Pahóm, “but will God let me live on it? I have lost my life, I have lost my life! I shall never reach that spot!” Pahóm looked at the sun, which had reached the earth: one side of it had already disappeared. With all his remaining strength he rushed on, bending his body forward so that his legs could hardly follow fast enough to keep him from falling. Just as he reached the hill it suddenly grew dark. He looked up -- the sun had already set!

He gave a cry: “All my labor has been for nothing,” thought he, and was about to stop, but he heard the Bashkírs still shouting, and remembered that though to him, from below, the sun seemed to have set, they on the hill could still see it. He took a long breath and ran up the hill. It was still light there. He reached the top and saw the soter that the chief had taken off from his head lying on the hillside. Before it sat the Chief laughing and holding his sides.

Again Pahóm remembered his dream, and he uttered a cry: his legs gave way beneath him, he fell forward and reached for the chief’s soter that was still lying on the ground.

“Ah, that’s a fine fellow!” exclaimed the Chief. “He has gained much land today!” Pahóm’s servant came running up and tried to raise him, but he saw that blood was coming from his mouth. Pahóm was dead!

The Bashkírs clicked their tongues to show their pity.

His servant picked up the taive and dug a grave long enough for Pahóm to lie in, and buried him in it. Six feet from his head to his heels was all he needed.

Appendix B: Reading Comprehension Questions

Chapter 1

- (1) What are the two sisters talking about during their conversation?
 - a. New types of food to try
 - b. Life in the city and in the country
- (2) What does Pahóm do for work?
 - a. He's a peasant farmer
 - b. He's a famous businessman
- (3) Who is listening to the women talking and Pahóm's talking to himself?
 - a. The children
 - b. The Devil

Chapter 2

- (1) What makes the peasants of the area upset?
 - a. The weather has been bad
 - b. The woman landowner is selling her property
- (2) Why does Pahóm complain to his wife?
 - a. He wants her to cook better food
 - b. He's worried about other people buy all of the land
- (3) Where do Pahóm and his wife get the money to buy the land?
 - a. They pay cash they already had
 - b. They sell things and borrow money

Chapter 3

- (1) Why does Pahóm get very angry?
 - a. Four of his animals die
 - b. Someone cuts down five of his trees
- (2) What do Pahóm and his visitor talk about?
 - a. A different community with good land
 - b. How they can trade animals
- (3) Why does Pahóm travel a long way?
 - a. To visit and look at a possible new home
 - b. To take a tour of the countryside

Chapter 4

- (1) What is the new community like?
 - a. It has better land and Pahóm has more animals and crops
 - b. It has a large river where Pahóm can catch fish
- (2) What does Pahóm begin to think is the problem?
 - a. He has to share the land, instead of having it for his own use
 - b. His family doesn't like the new community
- (3) What does the new traveler tell Pahóm?
 - a. He can get new farm equipment very cheaply
 - b. He can own a huge amount of land far away for a very low cost

Chapter 5

- (1) What do the Bashkírs live near to in their home area?
 - a. A different country
 - b. A river
- (2) In what way do the Bashkírs live?
 - a. They are happy, and the men are lazy
 - b. They are very hardworking and calm
- (3) Who do the people need to ask in order to give Pahóm land?
 - a. The chief of the Bashkírs
 - b. The council of leaders

Chapter 6

- (1) Why is Pahóm worried?
 - a. He wants his land to be given in writing
 - b. He wants to be able to buy something from the Bashkírs
- (2) Why is Pahóm surprised by the offer?
 - a. They want to give him all the land he can walk around in a day
 - b. They want to give him all the land he can see
- (3) What must Pahóm do to get the land?
 - a. Walk to the next town in time
 - b. Come back to the starting place in time

Chapter 7

- (1) What happens when Pahóm finally falls asleep?
 - a. He wakes up quickly afterward when an animal makes a sound
 - b. He dreams about someone laughing
- (2) Who is lying on the ground in Pahóm's dream?
 - a. The chief of the Bashkírs
 - b. Pahóm himself
- (3) What does Pahóm do when he wakes up?
 - a. Change his mind
 - b. Wake other people up

Chapter 8

- (1) What does the land look like?
 - a. It is rocky and hard
 - b. It is flat and beautiful
- (2) What makes Pahóm's walk difficult?
 - a. He doesn't like walking
 - b. It's very hot that day
- (3) What does Pahóm realize?
 - a. He needs to go straight back to the hill
 - b. He doesn't really like the land

Chapter 9

- (1) Why does Pahóm start to worry?
 - a. He thinks his family won't like the land
 - b. He's tired and doesn't think he can make it back in time
- (2) What are the Bashkírs doing on the hill?
 - a. Cheering for Pahóm
 - b. Playing games with each other
- (3) How much land does Pahóm need?
 - a. Six feet
 - b. As much as he can get

Appendix C: Characteristics of Target Pseudowords

Table 42

Lexical and Contextual Characteristics of Target Items in the Study

Original word(s) in text	Pseudoword replacement	Pseudoword Length (phonemes / letters/ syllables)	Pseudoword Orth neighbors	Spelling consistency (adapted from Chee et al., 2020): rime scores for each syllable to estimate spelling consistency	Mean concreteness rating (from Brysbaert et al., 2014) – max = 6	Mean SUBTLEX-US frequency count (from Brysbaert & New, 2009)	Nonword Bigram frequency mean	Mean (SD) for distance in text between instances (# of words)	Source of nonword
Businessman	Bancel <i>/bænsəl/</i>	6/6/2	1	Ban-: 0.7 -cel: 0.8 Total: 0.75	4.47	404	3586	515.7 (640.7)	Pellicer-Sánchez (2016)
Girdle / Clothes / Dress / Dressing Gown	Berrow <i>/beroʊ/</i>	5/6/2	3	Ber-: 1 -row: 0.7 Total: 0.85	4.73	1535	4381.6	668.7 (757.7)	Pellicer-Sánchez (2016)
Mutton / Bread	Cluff <i>/klʌf/</i>	4/5/1	3	-luff: 1 Total: 1.0	4.59	742.5	860.5	618.4 (293.8)	Elgort & Warren (2014)
Wheat / Rye	Creach <i>/kriʃ/</i>	4/6/1	2	-reach: 1 Total: 1.0	4.82	252.5	3826.4	665.5 (763.8)	Elgort & Warren (2014)
Steppe	Dault <i>/dɔlt/</i>	4/5/1	3	-ault: 1 Total: 1.0	4.6	5	1298.25	578.7 (493.8)	Elgort & Warren (2014)
Meadow / Pasture	Decop <i>/dəkɒp/</i>	5/5/2	2	Dec-: 0.4 -cop: 1 Total: 0.7	4.82	97	3372	501.7 (329.8)	Elgort & Warren (2014)
Village	Dello <i>/dɛləʊ/</i>	4/5/2	6	Del-: 0.4 -lo: 0.2 Total: 0.3	4.89	1712	3852.75	513 (413.6)	Elgort & Warren (2014)
Money	Emback <i>/embæk/</i>	5/6/2	2	Em-: 0.6 -back: 1.0 Total: 0.8	4.54	32679	1888.6	686.7 (395.6)	Elgort & Warren (2014)
Tent	Frine <i>/frɪn/</i>	4/5/1	3	-rine: 1.0 Total: 1.0	4.96	892	6680	508.1 (668.1)	Elgort & Warren (2014)
Cart	Gence <i>/dʒɛns/</i>	4/5/1	5	-ence: 1.0 Total: 1.0	4.89	461	3683.5	587.3 (836.3)	Elgort & Warren (2014)
Homestead / Home	Holter <i>/hɔltə/</i>	5/6/2	6	Hol-: 0.8 -ter: 1.0 Total: 0.9	4.56	32852.5	5505.6	517.7 (448.3)	Pellicer-Sánchez (2016)
Mile	Lastor <i>/lɑstə/</i>	6/6/2	2	Las: 0.4 -tor: 0.4 Total: 0.4	3.63	1071	4567.2	640 (677.4)	Elgort & Warren (2014)
Child/Children / Son(s)	Lidger <i>/lɪdʒə/</i>	4/6/2	3	Lid-: 0.9 -ger: 1.0 Total: 0.95	4.60	12639.7	4641	712.2 (349.2)	Elgort & Warren (2014)
Hillock	Merial <i>/mɛriəl/</i>	6/6/3	4	Mer-: 0.4 -i-: 0.1 -al: 1 Total: 0.5	4.93	1915	6182.8	633.4 (722.4)	Elgort & Warren (2014)

Town	Nurge /nʌrdʒ/	3/5/1	4	-urge: 1.0 Total: 1.0	4.64	12644	1587.75	516.3 (294.4)	Elgort & Warren (2014)
Rouble / Shilling (i.e., Dollar)	Pelon /pɛlɒn/	5/5/2	5	Pel-: 0.4 -on: 0.7 Total: 0.55	4.93	1410	4526.5	560.3 (515.8)	Elgort & Warren (2014)
Hole	Pleak /plɪk/	4/5/1	4	-leak: 0.5 Total: 0.5	4.81	2969	2951	646.2 (690.8)	Elgort & Warren (2014)
Cattle	Pogue /pɒʊg/	3/5/1	2	-ogue: 1.0 Total: 1.0	4.64	674	1242	554.7 (336.3)	Elgort & Warren (2014)
Tea	Punse /pʌns/	4/5/1	2	Pun-: 1.0 Total: 1.0	4.69	2990	3032.5	535.9 (627.5)	Elgort & Warren (2014)
Horse / Mare	Recibe /resɪb/	5/6/2	2	Rec-: 0.4 -ibe: 1.0 Total: 0.7	5	4737	3050.6	545.3 (480.7)	Elgort & Warren (2014)
Cow / Fox / Pig / Sheep	Shrick /ʃrɪk/	4/6/1	2	-rick: 0.6 Total: 0.6	4.88	2442.5	3151.2	685.8 (353)	Elgort & Warren (2014)
Kumiss (i.e., beer)	Snall /snɔl/	4/5/1	5	-all: 1.0 Total: 1.0	4.88	3850	3292.5	535.4 (693.4)	Elgort & Warren (2014)
Cap	Soter /sɔtə/	4/5/2	4	So-: 0.3 -ter: 1.0 Total: 0.65	4.59	956	6286	710.3 (588.3)	Pellicer- Sánchez (2016)
Steward / Innkeeper	Spiler /spɪlɪ/	5/6/2	1	Spi-: 0.3 -ler: 1.0 Total: 0.65	4.04	81	5246.8	689.4 (560.6)	Elgort (2017)
Spade / Sickle	Taive /teɪv/	3/5/1	2	-aive: 0.55 Total: 0.55	4.67	73	2584.5	715.5 (907.6)	Elgort & Warren (2014)

Appendix D: Form Recognition Test Items and Format

Targets:

Bancel, Berrow, Cluff, Creach, Dault, Decop, Dello, Emback, Frine, Gence, Holter, Lastor, Lidger, Merial, Nurge, Pelon, Pleak, Pogue, Punse, Recibe, Shrick, Snall, Soter, Spiler, Taive

Distractors:

From Webb (2008): Ancon, Cader, Dangy, Denent, Faddam, Hodet, Masco, Pacon, Sagod, Tasper

From Elgort & Warren (2014): Afuse, Extel, Grude, Flane, Staim, Trimp, Modium, Outlad, Seacon, Surmit, Alirn, Merly, Detise, Infate, Wockey

Format:

Randomized order of presentation in visual and auditory modalities using Psychopy (Peirce et al., 2019); for auditory items, participants were permitted to repeat the audio twice

Appendix E: Meaning Recall Test Items and Format

Instructions: write a translation of each word you see or hear in your home or first language. Then, click on how confident you are in that translation for each word.

1. _____

2. _____

3. _____

4. _____

5. _____

6. _____

7. _____

8. _____

9. _____

10. _____

11. _____

12. _____

13. _____

14. _____

15. _____

16. _____

17. _____

18. _____

19. _____

20. _____

21. _____

22. _____

23. _____

24. _____

25. _____

Format:

All items were target items; presented in randomized order and randomized modality using Psychopy (Peirce et al., 2019); for auditory items, participants were permitted to repeat the audio twice

Appendix F: Meaning Recognition Test Items and Format

Overall instructions: the following questions are about the meaning of words in the story. Do your best to remember the meaning of each word from the story, and choose one answer.

Stem	Answer + distractors (correct answer in bold)
A <u>gence</u> is...	A. a house B. a cart C. a person
A <u>dault</u> is...	A. a treat B. a walk C. a place
<u>Berrow</u> is...	A. something you wear B. something you eat C. something you look at
<u>Punse</u> is...	A. something to drink B. someone to speak with C. something to wear
A <u>merial</u> is...	A. a person B. a hill C. a home
<u>Creach</u> is...	A. something that grows B. something you eat C. something you wear
A <u>lastor</u> is...	A. a weight B. a distance C. a person
A <u>shrick</u> is...	A. a bird B. a person C. an animal
<u>Snall</u> is...	A. a person B. a drink C. a vehicle
<u>Emback</u> is...	A. money B. power C. family
A <u>soter</u> is...	A. an animal B. a hat C. a town
A <u>nurge</u> is...	A. a town B. a person C. something to wear
A <u>pelon</u> is...	A. clothing B. a person C. money
A <u>frine</u> is...	A. a tent B. a river C. a person

A <u>bancel</u> is...	A. a person B. a place C. an animal
A <u>spiler</u> is...	A. clothing B. a person C. a type of food
A <u>holter</u> is...	A. a place B. a person C. a tool
A <u>lidger</u> is...	A. clothing B. tools C. land
A <u>pleak</u> is...	A. a house B. an animal C. a hole
A <u>dello</u> is...	A. a home B. a village C. a nation
A <u>recibe</u> is...	A. an animal B. a place C. a type of food
A <u>taive</u> is...	A. a vehicle B. a tool C. a person
A <u>pogue</u> is...	A. a place to live B. an animal C. a river
A <u>decop</u> is...	A. a field B. a hat C. a person
<u>Cluff</u> is...	A. land B. people C. food

Format: presented in randomized order and randomized item modality using Psychopy (Peirce et al., 2019); for auditory items, participants were permitted to repeat the audio twice

Appendix G: Cloze Proficiency Measure

JD's cloze test "Man and his progress"

DIRECTIONS

1. Read the passage quickly to get the general meaning.
2. Write only one word in each blank next to the item number. Contractions are considered to be one word.
3. Check your answers.

EXAMPLE: The boy walked up the street. He stepped on a piece of ice.

He fell (1) down but he didn't hurt himself.

MAN AND HIS PROGRESS

Man is the only living creature that can make and use tools. He is the most teachable of living beings, earning the name of Homo sapiens. (1) ever restless brain has used the (2) and the wisdom of his ancestors (3) improve his way of life. Since (4) is able to walk and run (5) his feet, his hands have always (6) free to carry and to use (7). Man's hands have served him well (8) his life on earth. His development, (9) can be divided into three major (10), is marked by several different ways (11) life.

Up to 10,000 years ago, (12) human beings lived by hunting and (13). They also picked berries and fruits, (14) dug for various edible roots. Most (15), the men were the hunters, and (16) women acted as food gatherers. Since (17) women were busy with the children, (18) men handled the tools. In a (19) hand, a dead branch became a (20) to knock down fruit or (21) for tasty roots. Sometimes, an animal (22) served as a club, and a (23) piece of stone, fitting comfortably into (24) hand, could be used to break (25) or to throw at an animal. (26) stone was chipped against another until (27) had a sharp edge. The primitive (28) who first thought of putting a (29) stone at the end of a (30) made a brilliant discovery: he (31) joined two things to make a (32) useful tool, the spear. Flint, found (33) many rocks, became a common cutting (34) in the Paleolithic period of man's (35). Since no wood or bone tools (36) survived, we know of this man (37) his stone implements, with which he (38) kill animals, cut up the meat, (39) scrape the skins, as well as (40) pictures on the walls of the (41) where he lived during the winter.

(42) the warmer seasons, man wandered on (43) steppes of Europe without a fixed (44), always foraging for food. Perhaps the (45) carried nuts and berries in shells (46) skins or even in light, woven (47). Wherever they camped, the primitive people (48) fires by striking flint for sparks (49) using dried seeds, moss, and rotten (50) for tinder. With fires that he kindled himself, man could keep wild animals away and could cook those that he killed, as well as provide warmth and light for himself.

Answer keys

JD's cloze test "Man and his progress" - answer keys

- | Exact answer | Acceptable answer scoring would also include these possibilities |
|--------------|--|
| 1 His | our, man's the |
| 2 Knowledge | ideas, skill, work, teaching, wit, experience(s), talent, ingenuity, intelligence, cunning, culture, examples, mistakes, skills, words, thought, accomplishments, power, hands, nature, technique, instinct, will, information |
| 3 to | |

4	man	he
5	on	with, using, upon
6	been	hung, felt, remained
7	tools	freely, implements, readily, them, objects, carefully, productively, creatively, conventionally, weapons, adequately, diligently, efficiently, things
8	during	throughout, in, all, with, improving, for, through
9	which	however, often, also, since, that, conveniently, easily, historically, basically, thus
10	periods	groups, categories, parts, eras, stages, areas, sections, phases, topics, divisions, trends, steps, facets
11	of	for, towards, through, in
12	all	most, the, many, early, these, hungry, primitive, only
13	fishing	gathering, farming, killing, scrounging, scavenging, sleeping, trapping, foraging
14	and	often, some, the, ravenously
15	often	of, normally, always, trips, nights, important, times, emphatically
16	the	most, many, house, all, their, younger, older
17	the	most, many, often, all, married, these, primate, older
18	the	most, many, tough, constructive, primate, older, younger, all
19	man's	skilled, strong, learned, single, skillful, closed, big, empty, able, human('s), hunter's, person's, free, creative, right, needy, trained, deft, small, needed, coordinated
20	tool	club, pole, device, rod, stick, spear, instrument, weapon
21	dig	burrow, search, probe, excavate, test
22	bone	leg, horn, foot, tusk, tail, skull, had, arm, easily, hide
23	sharp	round, shaped, small, strong, chipped, fashioned, big, heavy, soft, rough, smooth, solid, sizeable, flat, thin, large, hard
24	the	one('s), man's, a, his
25	nuts	branches, wood, heads, bones, apart, trees, things, coconuts, down, bark, tinder, firewood, objects, food, sticks, shells, rocks, items, open, stone, ice, meat
26	one	the, softer, obsidian, shale, a, flat, hard, flint, glass, some, then, each, this
27	it	one, they, each
28	man	owner, being, person, human's, men, hunter, people, creature
29	sharp	small, sharpened, pointed, glass, lime, jagged, hard, large
30	stick	branch, log, rod, shaft, pole, bone, club
31	had	then, first, clumsily, tightly, tastefully, dexterously, cleverly, simply, double, securely, easily, soon, creatively, ingeniously, conveniently, would, suddenly, accidentally
32	very	portentiously, modern, useful, tremendously, necessarily, good, long, bad, quite, hunter's, extremely, intelligent, most, incredibly, new
33	in	that, among, by, using, inside, amongst, within, on, all
34	tool	stone, device, material, instrument, practice, utensil, implement, edge, piece, method, item, object
35	development	history, evolution, life, existence, time, discoveries, age, exploration, era, ancestry

36	have	actually, apparently, ever
37	by	and, used, from, through, for, using, had, made
38	could	would, did
39	and	or, then, carefully, would, help, skillfully
40	draw	carve, paint, create, the, hang, drawing, painting, place, sketch, engrave, some
41	cave(s)	place(s), animals, room
42	in	during, and, with
43	the	plain, unknown, to, flat, high, various, dry, toward, through, stone, across, aimless, barren, long, in, all, many
44	home	habitat, meal, income, weapons, diet, direction, destination, course, path, supplement, domain, place, camp, time, map, route, supply, lunch, plan, destiny, location, pattern, knowledge, foundation, appetite
45	women	men, man, primitives, wanderers, people, human, woman, children, voyager, group, families, hunter
46	or	and, with, of, animal, in, like, using, on, their, animal's, covered
47	baskets	bags, cloth(s), sacks, pouches, garments, material, fabric, chests, nets, hides, blankets, clothes
48	made	started, lit, built, lighted, used, produced, began
49	and	then, while, by, or, occasionally, together, also
50	wood	branches, bark, lumber, tree(s), skin, dung, roots, grass, timber, forage, leaves

First used and validated in

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *Modern Language Journal*, 64(3), 311-317.

Appendix H: LexTALE Proficiency Measure

Table 43

LexTALE Proficiency Measure Items and Correct Answers

Word / Nonword	Correct Answer
platory	No
denial	Yes
generic	Yes
mensible	No
scomful	Yes
stoutly	Yes
ablaze	Yes
kermshaw	No
moonlit	Yes
lofty	Yes
hurricane	Yes
flaw	Yes
alberation	No
unkempt	Yes
breeding	Yes
festivity	Yes
screech	Yes
savoury	Yes
plaudate	No
shin	Yes
fluid	Yes
spaunch	No
allied	Yes
slain	Yes
recipient	Yes
exprate	No
eloquence	Yes
cleanliness	Yes
dispatch	Yes
rebondicate	No
ingenious	Yes
bewitch	Yes
skave	No
plaintively	Yes
kilp	No
interfate	No
hasty	Yes

lengthy	Yes
fray	Yes
crumper	No
upkeep	Yes
majestic	Yes
magrity	No
nourishment	Yes
abergy	No
proom	No
turmoil	Yes
carbohydrate	Yes
scholar	Yes
turtle	Yes
fellick	No
destription	No
cylinder	Yes
ensorship	Yes
celestial	Yes
rascal	Yes
purrage	No
pulsh	No
muddy	Yes
quirty	No
pudour	No
listless	Yes
<u>wrought</u>	<u>Yes</u>

Appendix I: Reading Speed Proficiency Covariate

Screen 1 (61 words)

This story is about a man named Pahóm, who lives at a time long ago in Russia. Pahóm has many adventures in trying to find his way in life, and he works very hard to improve his and his family's situation. This story was written in the 19th century by Leo Tolstoy, a famous Russian writer. Press the "enter" key now.

Screen 2 (43 words)

If you like this story, you can find many others written by him and translated into English. As you read this story, think about the reading comprehension questions and the lessons we might learn from the story. Enjoy! Press the "enter" key now.

Appendix J: Language Background Questionnaire

Modified from LEAP-Q (Marian, Blumenfield, & Kaushanskaya, 2007)

Questions:

- (1) Participant number: _____
- (2) Age: _____
- (3) Sex: Female / Male
- (4) Country of origin: _____
- (5) English is your 1st / 2nd / 3rd / 4th / 5th language
- (6) What percentage of your time right now is spent using English? _____
- (7) How many years of formal education do you have? _____
- (8) Have you ever had one of the following problems: vision / hearing / language disability / learning disability? Tick all that apply, and explain any corrections.
- (9) The next several questions refer to your knowledge of English:
 - a. Age (in years) when you...
 - i. began acquiring English: _____
 - ii. became fluent in English: _____
 - iii. began reading in English: _____
 - iv. became fluent reading in English: _____
 - b. Please list the number of years and months you have spent in each language environment:
 - i. A country where English is spoken: _____
 - ii. A family where English is spoken: _____
 - iii. A school and/or working environment where English is spoken: _____
 - c. Please rate, on a scale of 1-10, to what extent you are currently exposed to English in the following contexts:
 - i. Interacting with friends:
1(never) 2(almost never) 3 4 5(half of the time) 6 7 8 9 10(always)
 - ii. Interacting with family:
1(never) 2(almost never) 3 4 5(half of the time) 6 7 8 9 10(always)
 - iii. Watching TV:
1(never) 2(almost never) 3 4 5(half of the time) 6 7 8 9 10(always)
 - iv. Listening to music:
1(never) 2(almost never) 3 4 5(half of the time) 6 7 8 9 10(always)
 - v. Reading:
1(never) 2(almost never) 3 4 5(half of the time) 6 7 8 9 10(always)

Appendix K: Debriefing Survey Questions

Debriefing questions for dissertation study

1. When you were reading the story, what were you trying to do? (both groups)
2. Did you notice the new words in the story? (both groups)
3. If you noticed them, did you try to learn the new words in the story? If so, how? (both groups)
4. Was it easy to understand the new words in the story? (both groups)
5. When you were reading, did you think that you would be tested on your knowledge of the new words? (both groups)
6. (RWL only) If reading with the audio helped you, in what way(s) did it help?
7. (RWL only) If reading with the audio made it more difficult to read, in what way(s) did it make it more difficult?
8. Is there anything else about your experience reading and learning the new words you'd like to express that you didn't already express on the previous questions? (both groups)

Appendix L: Model Comparisons for Inferential Analyses

Table 44

Model Comparisons for Each Eye-tracking Measure for Group Comparisons

Measure	Model	R^2 (Mar.)	R^2 (Cond.)	AIC	
Gaze Duration (Growth Curve)	Simplest	Log(gaze duration) ~ group + instance ¹ + group:instance ¹ + (1 Participant) + (1 Item)	0.015	0.14	40094
	Maximal ¹	Log(gaze duration) ~ group + instance ¹ + proficiency + PSTM + SRT+ Visit Count + group:instance ¹ + (1 Participant) + (1 Item)	0.064	0.152	39566
	Best fit	Log(gaze duration) ~ group + instance ¹ + proficiency + PSTM + Visit Count + group:instance ¹ + (1 Participant) + (1 Item)	0.073	0.174	39564
Total Reading Time (Growth Curve)	Simplest	Log(TRT) ~ group + instance ¹ + group:instance ¹ + (1 Participant) + (1 Item)	0.018	0.103	48034
	Maximal ¹	Log(TRT) ~ group + instance ¹ + proficiency + PSTM + SRT+ Visit Count + group:instance ¹ + (1 Participant) + (1 Item)	0.167	0.228	43406
	Best fit	Log(TRT) ~ group + instance ¹ + proficiency + PSTM + Visit Count + group:instance ¹ + (1 Participant) + (1 Item)	0.221	0.288	43405
Visit Count (Negative Binomial Logistic Regression)	Simplest	Visit Count ~ group + instance + group:instance + (1 Participant) + (1 Item)	0.004	0.054	79529
	Maximal ¹	N/A – model could not converge with all covariates	N/A	N/A	N/A
	Best fit	Visit Count ~ group + instance + proficiency + group:instance + (1 Participant) + (1 Item)	0.011	0.052	79510

¹Maximal models here do not include random slopes. None contributed to model fit, or resulted in model non-convergence, so Maximal here is defined as the maximal models including only random intercepts.

Table 45*Model Comparisons for Each Vocabulary Learning Outcome*

Measure	Model	R^2 (Mar.)	R^2 (Cond.)	AIC	
Form Recognition	Simplest	accuracy ~ group + item modality + group:item modality + (1 Participant) + (1 Item)	0.011	0.243	3183
	Maximal ¹	accuracy ~ group + item modality + proficiency + PSTM + SRT + TRT + group:item modality + (1 Participant) + (1 Item)	0.047	0.244	3172.2
	Best fit	accuracy ~ group + item modality + proficiency + group:item modality + (1 Participant) + (1 Item)	0.045	0.244	3167.3
Meaning Recognition	Simplest	accuracy ~ group + item modality + group:item modality + (1 Participant) + (1 Item)	0.004	0.237	3725.4
	Maximal ¹	accuracy ~ group + item modality + proficiency + PSTM + SRT + TRT + group:item modality + (1 Participant) + (1 Item)	0.033	0.241	3703.9
	Best fit	accuracy ~ group + item modality + proficiency + group:item modality + (1 Participant) + (1 Item)	0.032	0.237	3698.5
Meaning Recall	Simplest	accuracy ~ group + item modality + group:item modality + (1 Participant) + (1 Item)	0.043	0.51	1870.6
	Maximal ¹	accuracy ~ group + item modality + proficiency + PSTM + SRT + TRT + group:item modality + (1 Participant) + (1 Item)	0.137	0.503	1851.8
	Best fit	accuracy ~ group + item modality + proficiency + group:item modality + (1 Participant) + (1 Item)	0.131	0.503	1850.2

¹Maximal models here do not include random slopes. None contributed to model fit, or resulted in model non-convergence, so Maximal here is defined as the maximal models including only random intercepts.

Table 46*Model Comparisons for Readahead Variable Regressed on Outcomes within RWL Group*

Measure	Model	R^2 (Mar.)	R^2 (Cond.)	AIC	
Summed Total Reading Time	Simplest	TRT_summed ~ readahead + (1 Participant) + (1 Item)	0.009	0.569	74019
	Maximal ¹	TRT_summed ~ readahead + proficiency + PSTM + SRT + (1 Participant) + (1 Item)	0.054	0.574	74014
	Best fit	TRT_summed ~ readahead + proficiency + PSTM + (1 Participant) + (1 Item)	0.047	0.572	74014
Form Recognition	Simplest	accuracy ~ readahead + (1 Participant) + (1 Item)	0.003	0.236	1559
	Maximal ¹	accuracy ~ readahead + proficiency + PSTM + SRT + TRT + (1 Participant) + (1 Item)	0.044	0.242	1556.5
	Best fit	accuracy ~ readahead + proficiency + TRT + (1 Participant) + (1 Item)	0.04	0.24	1553.7
Meaning Recognition	Simplest	accuracy ~ readahead + (1 Participant) + (1 Item)	0.004	0.262	1833
	Maximal ¹	accuracy ~ readahead + proficiency + PSTM + SRT + TRT + (1 Participant) + (1 Item)	0.043	0.266	1823
	Best fit	accuracy ~ readahead + proficiency + TRT + (1 Participant) + (1 Item)	0.043	0.266	1819.3
Meaning Recall	Simplest	accuracy ~ readahead + (1 Participant) + (1 Item)	0.015	0.449	1125
	Maximal ¹	accuracy ~ readahead + proficiency + PSTM + SRT + TRT + (1 Participant) + (1 Item)	0.139	0.458	1115.9
	Best fit	accuracy ~ readahead + proficiency + TRT + (1 Participant) + (1 Item)	0.107	0.456	1117.9

¹Maximal models here do not include random slopes. None contributed to model fit, or resulted in model non-convergence, so Maximal here is defined as the maximal models including only random intercepts.

References

- Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*, 814 – 823. <http://dx.doi.org/10.1111/j.1467-9280.2006.01787.x>
- Adesope, O., & Nesbit, J. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology*, *104*(1), 250.
- Akbulut, Y. (2007). Effects of multimedia annotations on incidental vocabulary learning and reading comprehension of advanced learners of English as a foreign language. *Instructional Science*, *35*(6), 499-517.
- Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134-142.
- Antúnez, M., Milligan, S., Hernández-Cabrera, J. A., Barber, H. A., & Schotter, E. R. (2022). Semantic parafoveal processing in natural reading: Insight from fixation-related potentials & eye movements. *Psychophysiology*, *59*(4), e13986.
- Audacity Team (2023). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 3.0.0 retrieved June 6, 2022 from <https://audacityteam.org/>
- Baddeley, A. (1998). Recent developments in working memory. *Current opinion in neurobiology*, *8*(2), 234-238.
- Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445-459.
- Barcroft, J. (2009). Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Quarterly*, *43*(1), 79-103.

- Barcroft, J. (2015). Can retrieval opportunities increase vocabulary learning during reading?. *Foreign Language Annals*, 48(2), 236-249.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Birch, S., & Rayner, K. (2010). Effects of syntactic prominence on eye movements during reading. *Memory & Cognition*, 38(6), 740-752.
- Bisson, M., van Heuven, W., Conklin, K., & Tunney, R. (2013). Incidental acquisition of foreign language vocabulary through brief multi-modal exposure. *PLoS One*, 8(4), e60912.
- Bisson, M., Van Heuven, W., Conklin, K., & Tunney, R. (2014). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language learning*, 64(4), 855-877.
- Bisson, M., Van Heuven, W., Conklin, K., & Tunney, R. (2015). The role of verbal and pictorial information in multimodal incidental acquisition of foreign language vocabulary. *Quarterly Journal of Experimental Psychology*, 68(7), 1306-1326.
- Bley-Vroman, R. (1988). The fundamental character of foreign language learning. In Rutherford, W. & Sharwood Smith, M. (Eds.), *Grammar and second language teaching* (pp. 19–30). Rowley, MA: Newbury House.
- Blum, I., Koskinen, P., Tennant, N., Parker, E., Straub, M., & Curry, C. (1995). Using audiotaped books to extend classroom literacy instruction into the homes of second-language learners. *Journal of Reading Behavior*, 27(4), 535-563.
- Bonilla, C., Golonka, E., Pandža, N., Linck, J., Michael, E., Clark, M., Lancaster, A., & Richardson, D. (2020). Leveraging Spanish knowledge and cognitive aptitude in

- Portuguese learning. *Linguistic Approaches to Portuguese as an Additional Language*, 24, 191-230.
- Bordag, D., Kirschenbaum, A., Tschirner, E., & Opitz, A. (2015). Incidental acquisition of new words during reading in L2: Inference of meaning and its integration in the L2 mental lexicon. *Bilingualism: Language and Cognition*, 18(3), 372-390.
- Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition*, 25(2), 185-201.
- Borro, I. (2021). *Enhanced incidental learning of formulaic sequences by Chinese learners of Italian* (Doctoral Dissertation, University of Portsmouth).
- Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). American Council on Education: Praeger.
- Brown, J.D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal*, 64(3), 311-317.
- Brown, D. (2021). Incidental vocabulary learning in a Japanese university L2-English language classroom over a semester. *TESOL Journal*, 12(4), e595.
- Brown, R., Waring, B., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language* 20(2), 136-163.
- Bruton, A., García López, M., & Esquiliche Mesa, R. (2011). Incidental vocabulary learning: an impracticable term? *TESOL Quarterly*, 45, 4, 759-768.
- Brysbaert, M. (2022). Word Recognition II: Phonological Coding in Reading. *The Science of Reading: A Handbook*, 79-101.

- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977-990.
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: a tutorial. *Journal of Cognition*, *1*(1), 1-20. <https://doi.org/10.5334/joc.10>
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, *46*, 904-911.
- Bunting, M., Cowan, N., & Scott Saults, J. (2006). How does running memory span work? *Quarterly Journal of experimental psychology*, *59*(10), 1691-1700.
- Bürki, A. (2010). Lexis that rings a bell: on the influence of auditory support in vocabulary acquisition. *International Journal of Applied Linguistics*, *20*(2), 206-231.
- Carbo, M. (1978). Teaching reading with talking books. *The Reading Teacher*, *32*(3), 267-273.
- Carney, N. (2021). Diagnosing L2 listeners' difficulty comprehending known lexis. *TESOL Quarterly*, *55*(2), 536-567.
- Carrol, G., & Conklin, K. (2020). Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences. *Language and Speech*, *63*(1), 95-122.
- Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095-1102.
- Chang, A. C. S., & Millett, S. (2015). Improving reading rates and comprehension through audio-assisted extensive reading for beginner learners. *System*, *52*, 91-102.

- Chaudron, C. (1985). Intake: on models and methods for discovering learners' processing of input. *Studies in Second Language Acquisition*, 7(1), 1-14.
- Chee, Q. W., Chow, K. J., Yap, M. J., & Goh, W. D. (2020). Consistency norms for 37,677 English words. *Behavior Research Methods*, 52(6), 2535-2555.
- Chen, Y. (2021). Comparing incidental vocabulary learning from reading-only and Reading-while-Listening. *System*, 97, 102442.
- Chen, X., Dong, Y., & Yu, X. (2018). On the predictive validity of various corpus-based frequency norms in L2 English lexical processing. *Behavior Research Methods*, 50, 1–25. <http://dx.doi.org/10.3758/s13428-017-1001-8>
- Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, 31(5), 693-713.
- Chun, D. (2016). The role of technology in SLA research. *Language Learning & Technology*, 20(2), 98-115.
- Chung, H. (1995). Effects of elaborative modification on second language reading comprehension and incidental vocabulary learning. *University of Hawai'i Working Papers in ESL*, 14(1), 27-61.
- Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be conscious. *Frontiers in Psychology*, 2. <https://doi.org/10.3389/fpsyg.2011.00086>
- Cobb, T. (2022). Compleat Web VP v2.6 [computer program]. Accessed May 18, 2022 at <https://www.lexutor.ca/vp/comp/>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, 65(3), 145.

- Coltheart, M. (2005). Modeling reading: the dual-route approach. In Snowling & Hulme (Eds.), *The Science of Reading: A Handbook*, 6-23. Malden, MA: Blackwell.
- Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press.
- Conklin, K., Alotaibi, S., Pellicer-Sánchez, A., & Vilkaitė-Lozdienė, L. (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research*, 36(3), 257-276.
- Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, 32, 45-61.
- Conklin, K., & Alotaibi, S. (2023). Eye-tracking reading-while-listening: challenges and methodological considerations in vocabulary research. *Research Methods in Applied Linguistics*, 2(3), 100086. <https://doi.org/10.1016/j.rmal.2023.100086>
- Cook, S., Pandža, N., Lancaster, A., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, 7, 1-17.
- Cop, U., Drieghe, D., & Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS one*, 10(8), e0134008.
- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28(3), 369-382.
- Darcy, I. (2022). From fuzzy to fine-grained representations in the developing lexicon. *Bilingualism: Language and Cognition*, 25(2), 206-207.

- Darcy, I., & Thomas, T. (2019). When blue is a disyllabic word: Perceptual epenthesis in the mental lexicon of second language learners. *Bilingualism: Language and Cognition*, 22(5), 1141-1159.
- DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22(4), 499-533.
- DeKeyser, R.M. (2008). Implicit and explicit learning. In Long, M., & Doughty, C. (Eds.), *The Handbook of Second Language Acquisition*, 312-348.
- Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, 79(2), 162.
- Dewaele, J. M. (2009). Individual differences in second language acquisition. In Ritchie, W., & Bhatia, T. (Eds.). *The New Handbook of Second Language Acquisition* (Brill), 623-646.
- Diao, Y., & Sweller, J. (2007). Redundancy in foreign language reading comprehension instruction: concurrent written and spoken presentations. *Learning and Instruction*, 17(1), 78–88.
- Dörnyei, Z. (2006). Individual differences in second language acquisition. *AILA review*, 19(1), 42-68.
- Doughty, C. (2008). Instructed SLA: constraints, compensation, and enhancement. In Long, M., & Doughty, C. (Eds.), *The Handbook of Second Language Acquisition*, 256-310.
- Doughty, C. J., & Mackey, A. (2021). Language aptitude: Multiple perspectives. *Annual Review of Applied Linguistics*, 41, 1-5. doi:10.1017/S0267190521000076
- Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In Doughty, C., & Williams, J. (Eds.). *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press, 1998. p. 197-261.

- Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, 26(2), 163-188.
- Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61(2), 367-413.
- Elgort, I. (2017). Incorrect inferences and contextual word learning in English as a second language. *Journal of the European Second Language Association*, 1(1), 1-11, DOI: <https://doi.org/10.22599/jesla.3>
- Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning*, 64(2), 365-414.
- Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018a). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 40(2), 341-366.
- Elgort, I., Candry, S., Boutorwick, T. J., Eyckmans, J., & Brysbaert, M. (2018b). Contextual word learning with form-focused and meaning-focused elaboration. *Applied Linguistics*, 39(5), 646-667.
- Ellis, N. (1994). Implicit and explicit language learning—An overview. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 1-31). New York: Academic Press.
- Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in second language acquisition*, 27(2), 141-172.
- Ellis, N. C. (2008). Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In *Handbook of cognitive linguistics and second language acquisition* (pp. 382-415). Routledge.

- File, K. A., & Adams, R. (2010). Should vocabulary instruction be integrated or isolated? *TESOL Quarterly*, 44(2), 222-249.
- Flege, J., & Bohn, O.S. (2021). The revised speech learning model (SLM-r). In Wayland, R. (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress*, 3-83. Cambridge University Press.
- Foster, J.L., Shipstead, Z., Harrison, T.L., Hicks, K.L., Redick, T.S., & Engle, R.W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, 43(2), 226-236.
- Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1), 104-115.
- Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology Section A*, 54(1), 1-30.
- Godfroid, A. (2020a). Sensitive measures of vocabulary knowledge and processing: expanding Nation's framework. In Webb, S. (Ed.), *The Routledge Handbook of Vocabulary Studies* (New York: Taylor & Francis), 433-453.
- Godfroid, A. (2020b). *Eye tracking in second language acquisition and bilingualism: a research synthesis and methodological guide*. New York: Routledge.
- Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: gauging the role of attention in L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, 35(3), 483-517.

- Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon, H.J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, 21(3), 563-584.
- Godfroid, A., & Hui, B. (2020). Five common pitfalls in eye-tracking research. *Second Language Research*, 36(3), 277–305. <https://doi.org/10.1177/0267658320921218>
- Goldman-Eisler, F. (1961). The significance of changes in the rate of articulation. *Language and Speech* 4, 171–74.
- Gor, K., & Cook, S. (2020). A mare in a pub? Nonnative facilitation in phonological priming. *Second Language Research*, 36(1), 123-140.
- Gor, K., Cook, S., Bordag, D., Chrabaszcz, A., & Opitz, A. (2021). Fuzzy lexical representations in adult second language speakers. *Frontiers in Psychology*, 12:732030.
- Goswami, U., & Bryant, P. (1990). Essays in developmental psychology series. *Phonological Skills and Learning to Read*. New Jersey: Lawrence Erlbaum.
- Grainger, J., & Ziegler, J. C. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology*, 2, 54.
- Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second language research*, 29(3), 311-343.
- Griffiths, R. (1990). Speech rate and NNS comprehension: A preliminary study in time-benefit analysis. *Language Learning*, 40(3), 311-336.
- Hamrick, P., & Pandža, N. B. (2020). Contributions of semantic and contextual diversity to the word frequency effect in L2 lexical access. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 74(1), 25.

- Harm, M., & Seidenberg, M. (2004). Computing the meanings of words in reading: division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662–720.
- Hastie, T.J., & Tibsharani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, *1*, 336-337.
- Hatami, S. (2017). The impact of learner-related variables on second language incidental vocabulary acquisition through listening. *Vocabulary Learning and Instruction*, *1*, 1-20.
- Hazenbergh, S., & Hulstijn, J. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied linguistics*, *17*(2), 145-163.
- Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207-223.
- Hui, B. (2021). *A construct validation study of implicit and time sensitive vocabulary measures*. Michigan State University (doctoral dissertation).
- Hui, B. (2024). Scaffolding comprehension with reading while listening and the role of reading speed and text complexity. *The Modern Language Journal*, Early View. doi: <https://doi.org/10.1111/modl.12905>
- Hui, B. & Godfroid, A. (in-principle acceptance). Audiobooks decomposed: Toward a psycholinguistic account of the benefits of reading-while-listening for verbal comprehension. *Language Learning*.
- Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), 1089-1115.
- Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: A

- reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 258-286). Cambridge University Press.
- Hulstijn, J., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign students: the influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*(3), 327-339.
- Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language learning*, *51*(3), 539-558.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434-446.
- Jelani, N., & Boers, F. (2018). Examining incidental vocabulary acquisition from captioned video: Does test modality matter? *ITL-International Journal of Applied Linguistics*, *169*(1), 169-190.
- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: a meta-analysis. *Language Learning*, *64*(1), 160–212. <https://doi.org/10.1111/lang.12034>
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, *21*(1), 47-77.
- Jiang, N. (2021). Examining L1 influence in L2 word recognition: A case for case. *Journal of Second Language Studies*, *4*(1), 1-18.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601-625.
- Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human factors*, *40*(1), 1-17.

- Kaushanskaya, M. (2012). Cognitive mechanisms of word learning in bilingual and monolingual adults: The role of phonological memory. *Bilingualism: Language and Cognition*, 15(3), 470-489.
- Kim, Y. (2006). Effects of input elaboration on vocabulary acquisition through reading by Korean learners of English as a foreign language. *TESOL Quarterly*, 40(2), 341-373.
- Kim, Y. (2011). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language learning*, 61, 100-140.
- Kinchla, R.A. (1974). Detecting target elements in multielement arrays: a confusability model. *Perception and Psychophysics*, 15(1), 149-158.
- Kobayashi Hillman, K. (2020). Effects of different types of auditory input on incidental vocabulary learning by L2 Japanese speakers. Unpublished dissertation.
- Krashen, S. D. (1981). Bilingual education and second language acquisition theory. *Schooling and language minority students: A theoretical framework*, 51-79.
- Kuder, G.F. & Richardson, M.W. (1937) The Theory of the Estimation of Test Reliability. *Psychometrika*, 2, 151-160. <https://doi.org/10.1007/BF02288391>
- Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Lõo, K., Marelli, M., ... Usal, K. A. (2023). Text reading in English as a second language: evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*, 45(1), 3–37. <https://doi.org/10.1017/S0272263121000954>
- Laufer, B. (2001). Reading, word-focused activities and incidental vocabulary acquisition in a second language. *Prospect*, 16(3), 44-54.

- Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review*, 59(4), 567-587.
- Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 15(3), 209-225.
- Lee, S. K., & Huang, H. T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, 30(3), 307-331.
- Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid test for advanced learners of English. *Behavioral Research Methods*, 44(2), 325-343.
- Leung, J., & Williams, J. The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition*, 33(1), 33-55.
- Lewandowski, L., & Kobus, D. (1993). The effects of redundancy in bimodal word processing. *Human Performance*, 6(3), 229-239.
- Li, M., Jiang, N., & Gor, K. (2017). L1 and L2 processing of compound words: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition*, 20(2), 384-402.
- Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, 36(5), 1247-1282.
- Linck, J., Hughes, M., Campbell, S., Silbert, N., Tare, M., Jackson, S., Smith, B., Bunting, M., & Doughty, C. (2013). Hi-LAB: a new measure of aptitude for high-level language proficiency. *Language Learning*, 63(3), 530-566.

- Llompert, M., & Reinisch, E. (2019). Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts. *Bilingualism: Language and Cognition*, 22(5), 1085-1100.
- Loewen, S. (2014). The acquisition of vocabulary. In *Introduction to Instructed Second Language Acquisition* (pp. 107-126). Routledge.
- Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *The Modern Language Journal*, 105(1), 187-193.
- Long, M. H. (1980). Inside the “black box”: Methodological issues in classroom research on language learning. *Language learning*, 30(1), 1-42.
- Long, M. H. (1981). Questions in foreigner talk discourse. *Language learning*, 31(1), 135-157.
- Long, M. H. (1990). The least a second language acquisition theory needs to explain. *TESOL Quarterly*, 24(4), 649-666.
- Long, M. M. (2017). Instructed second language acquisition (ISLA): geopolitics, methodological issues, and some major research questions. *Instructed Second Language Acquisition*, 1(1), 7-44.
- Maie, R., & DeKeyser, R. M. (2020). Conflicting evidence of explicit and implicit knowledge from objective and subjective measures. *Studies in Second Language Acquisition*, 42(2), 359-382.
- Malone, J. (2018). Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, 40(3), 651–675.

- Marian, V., Blumenfeld, H., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940-967.
- Martin, K. & Ellis, N.C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition, 34*(3) , 379-413.
- Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187.
- Mayer, R. E., & Fiorella, L. (2014). 12 principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In *The Cambridge handbook of multimedia learning* (Vol. 279). New York, NY: Cambridge University Press.
- Mayer, R. E., & Johnson, C. I. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology, 100*(2), 380.
- McKoon, G., Ratcliff, R., Ward, G., & Sproat, R. (1993). Syntactic prominence effects on discourse processes. *Journal of Memory and Language, 32*(5), 593-607.
- McQuillan, J. (2019). Where do we get our academic vocabulary? Comparing the efficiency of direct instruction and free voluntary reading. *The Reading Matrix: An International Online Journal, 19*(1), 129-138.
- McQuillan, J., & Krashen, S. D. (2008). Commentary: Can free reading take you all the way? A response to Cobb (2007). *Language Learning & Technology, 12*(1), 104-108.

- McZgee, V. E., & Carleton, W. T. (1970). Piecewise regression. *Journal of the American Statistical Association*, 65(331), 1109-1124.
- Mirman, D. (2014). Growth curve analysis: A hands-on tutorial on using multilevel regression to analyze time course data. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).
- Mohamed, A. A. (2018). Exposure frequency in L2 reading: An eye-movement perspective of incidental vocabulary learning. *Studies in Second Language Acquisition*, 40(2), 269–293. <https://doi.org/10.1017/S0272263117000092>
- Montali, J., & Lewandowski, L. (1996). Bimodal reading: benefits of a talking computer for average and less-skilled readers. *Journal of Learning Disabilities*, 29(3), 271-279.
- Montero Perez, M. (2020). Incidental vocabulary learning through viewing video: The role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, 42(4), 749-773.
- Moravcsik, J. E., & Healy, A. F. (1998). Effect of syntactic role and syntactic prominence on letter detection. *Psychonomic Bulletin & Review*, 5(1), 96-100.
- Moreno, R., & Mayer, R. E. (2002). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology*, 94(3), 598.
- Muñoz, C., Pujadas, G., & Pattemore, A. (2021). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research*, Online First.
- Nagy, W. E., Herman, P.A., & Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233–53.
- Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge, England: Cambridge University Press.

- Nation, I. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59-82.
- Nation, I.S.P. (2017). The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. Available from <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>
- Nation, P., & Meara, P. (2013). 3 | Vocabulary. In *An Introduction to Applied Linguistics* (pp. 44-62). Routledge.
- Nation, I.S.P., & Chung, M. (2009). Teaching and testing vocabulary. In Long, M.H., & Doughty, C.J. (eds.), *Handbook of language teaching* (pp. 543-559). Malden, MA: Wiley-Blackwell.
- Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal*, 105(1), 218-236.
- Nguyen, C. D., & Boers, F. (2019). The effect of content retelling on vocabulary uptake from a TED talk. *TESOL Quarterly*, 53(1), 5-29.
- Paivio, A. (2013). Dual coding theory, word abstractness, and emotion: a critical review of Kousta et al. (2011). *Journal of Experimental Psychology: General*, 142(1), 282-287. doi: 10.1037/a0027004
- Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: imagery or dual coding?. *Cognitive Psychology*, 5(2), 176–206. doi:10.1016/0010-0285(73)90032-7
- Paribakht, T. S., & Wesche, M. (1999). Reading and “incidental” L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition*, 21(2), 195-224.
- Pawlas, A.A., Ramig, L.O., & Countryman, S. (1996). Perceptual Voice and Speech Characteristics in Patients with Idiopathic Parkinson’s Disease. *NCVS Status Report #10*,

- 79-87. Available at
<http://www.ncvs.org/ProgressReports/NCVS%20Status%20&%20Progress%20Report%20Vol.%2010,%20Nov%201996%20copy.pdf>
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203.
- Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading. *Studies in Second Language Acquisition*, *38*(1), 97-130.
- Pellicer-Sánchez, A., & Boers, F. (2018). Pedagogical approaches to the teaching and learning of formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 153–173). Routledge.
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383.
- Montero Pérez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal*, *99*(2), 308-328.
- Ota, M., Hartsuiker, R. J., & Haywood, S. L. (2009). The KEY to the ROCK: Near-homophony in nonnative visual word recognition. *Cognition*, *111*(2), 263-269.
- Ota, M., Hartsuiker, R. J., & Haywood, S. L. (2010). Is a FAN always FUN? Phonological and orthographic effects in bilingual visual word recognition. *Language and speech*, *53*(3), 383-403.
- Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *40*(3), 551-577.

- Pitts, M., White, H., & Krashen, S. (1989). Acquiring the second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language*, 5(2), 271-275.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878-912.
- Pollack, I., Johnson, I. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology*, 57, 137–146.
- Pollatsek, A., Reichle, E., and Rayner, K. Tests of the EZ Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive psychology* 52(1), 1-56.
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language learning*, 52(3), 513-536.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372.
- Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr., C. (2012). *Psychology of reading*. Psychology Press.
- Rayner, K., & Pollatsek, A. (2016). Eye movements in reading a tutorial review. *Attention and performance XII*, 327-362.
- Rebuschat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, 33(4), 829-856.
- Ren, J., & Wang, M. (2023). Sensitivity to word endings as probabilistic orthographic cues to lexical stress among English as second language learners. *Memory & Cognition*, 51, 1881-1897.

- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4), 445-476.
- Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic bulletin & review*, 16, 1-21.
- Reichle, E. D., Liversedge, S. P., Drieghe, D., Blythe, H. I., Joseph, H. S., White, S. J., & Rayner, K. (2013). Using EZ Reader to examine the concurrent development of eye-movement control and reading skill. *Developmental Review*, 33(2), 110-149.
- Robinson, P. (2002). Learning conditions, aptitude complexes, and SLA. In Robinson, P. (Ed.), *Individual Differences and Instructed Language Learning*, 113-133. Philadelphia: John Benjamins.
- Robinson, P. (2008). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 631–678). Malden, MA: Blackwell.
- Robinson, P. (2012). Individual differences, aptitude complexes, SLA processes, and aptitude test development. In *New Perspectives on Individual Differences in Language Learning and Teaching* (pp. 57-75). Springer, Berlin, Heidelberg.
- Rodgers, M. P., & Webb, S. (2020). Incidental vocabulary learning through viewing television. *ITL-International Journal of Applied Linguistics*, 171(2), 191-220.
- Roehr, K (2012) Aptitude treatment interaction (ATI) research. In Robinson, P, (ed.) *The Routledge Encyclopedia of Second Language Acquisition*. Routledge, 31–35. New York: Routledge.

- Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition through reading. *Studies in Second Language Acquisition*, 21(4), 589-619.
- Rott, S. (2007). The effect of frequency of input-enhancements on word learning and text comprehension. *Language Learning*, 57(2), 165-199.
- Sánchez Gutiérrez, C. H., Serrano, M. P., & García, P. R. (2019). The effects of word frequency and typographical enhancement on incidental vocabulary learning in reading. *Journal of Spanish Language Teaching*, 6(1), 14-31.
- Saragi, T. (1978). Vocabulary learning and reading. *System*, 6(2), 72-8.
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158.
- Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *Consciousness in second language learning*, 11, 237-326.
- Schmidt, R. (2012). Attention, awareness, and individual differences in language learning. *Perspectives on Individual Characteristics and Foreign Language Education*, 6, 27.
- Schmitt, N. (2000). Key concepts in ELT. *ELT Journal*, 54(4), 400-401.
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.
- Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47(4), 484-503.
- Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics*, 74, 5-35.

- Segalowitz, N. (1997). Individual differences in second language acquisition. In de Groot, A., & Kroll, J. (Eds.), *Tutorials in Bilingualism: Psycholinguistic Perspectives*, 85-112. Mahwah, NJ: Erlbaum.
- Segalowitz, N. (2008). Automaticity and second languages. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 382–406). Malden, MA: Blackwell.
- Seol, H. (2020). *snowIRT: Item Response Theory for jamovi*. [jamovi module]. Retrieved from <https://github.com/hyunsooseol/snowIRT>.
- Serrano, R., & Pellicer-Sánchez, A. (2019). Young L2 learners' online processing of information in a graded reader during reading-only and reading-while-listening conditions: A study of eye-movements. *Applied Linguistics Review*, 13(1), 49-70.
- Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition*, 13(2), 275-298.
- Spit, S., Andringa, S., Rispens, J., & Aboh, E. O. (2021). Do kindergarteners develop awareness of the statistical regularities they acquire? *Language Learning*, 71(2), 573–611. <https://doi.org/10.1111/lang.12445>
- SR Research (2010) EyeLink 1000 User Manual (version 1.5.0). Available at <http://srresearch.jp/support/EyeLink%201000%20User%20Manual%201.5.0.pdf> (accessed 1 June 2022).
- Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.
- Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21(4), 360-407.

- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.
- Stekić, K., Ilić, O., Ković, V., & Savić, A. M. (2023). ERP Indicators of Phonological Awareness Development in Children: A Systematic Review. *Brain Sciences*, *13*(2), 290.
- Strauss, A., & Corbin, J. M. (1997). *Grounded theory in practice*. Sage.
- Suk, N. (2017). The effects of extensive reading on reading comprehension, reading rate, and vocabulary acquisition. *Reading Research Quarterly*, *52*(1), 73-89.
- Suzuki, Y. (2015). Using new measures of implicit L2 knowledge to study the interface of explicit and implicit knowledge (Doctoral dissertation, University of Maryland, College Park).
- Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*(5), 1229-1261.
- Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: insights from individual differences in cognitive aptitudes. *Language Learning*, *67*(4), 747-790.
- Sweet, H. 1899. *The Practical Study of Languages. A Guide for Teachers and Learners*. London: Dent.
- Sweller, J. (1988). Cognitive load during problem solving: Effects of learning. *Cognitive Science*, *12*, 257-285.
- Sweller, J. (2011). Cognitive load theory. In Ross, B. (Ed.), *Psychology of Learning and Motivation* (pp. 37-76). Academic Press.

- Syodorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology, 14*(2), 50-73.
- Taguchi, E., Gorsuch, G., Lems, K., & Rosszell, R. (2016). Scaffolding in L2 reading: how repetition and an auditory model help readers. *Reading in a Foreign Language, 28*(1), 101-117.
- Teng, F. (2018). Incidental vocabulary acquisition from reading-only and reading-while-listening: A multi-dimensional approach. *Innovation in Language Learning and Teaching, 12*(3), 274-288.
- Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition, 16*(2), 183-203.
- Tragant, E., & Vallbona, A. (2018). Reading while listening to learn: young EFL learners' perceptions. *ELT Journal, 72*(4), 395-404.
- Tragant Mestres, E., & Pellicer-Sánchez, A. (2019). Young EFL learners' processing of multimodal input: examining learners' eye movements. *System, 80*, 212-223.
- Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language, 28*(2), 127-154.
- Tuzcu, A. (2023). *Unimodal and bimodal input in incidental vocabulary learning: cognitive processes and the development of different knowledge types*. (Doctoral Dissertation, Michigan State University).
- Uchihara, T., & Clenton, J. (2023). The role of spoken vocabulary knowledge in second language speaking proficiency. *The Language Learning Journal, 51*(3), 376-393.

- Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning, 69*(3), 559-599.
- Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). Frequency of exposure influences accentedness and comprehensibility in learners' pronunciation of second language words. *Language Learning, FirstView*.
- Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*(1-2), 231-270.
- Ullman, M. T. (2006). The declarative/procedural model and the shallow structure hypothesis. *Applied Psycholinguistics, 27*(1), 97-105.
- Vafaei, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition, 42*(2), 383-410.
- Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology, 88*(11), 2766-2772.
- Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219-258.
- Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect size-driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition, 44*(5), 1424-1448.
- Wang, M., Koda, K., & Perfetti, C. (2003). Alphabetic and nonalphabetic L1 effects in English word identification: a comparison of Korean and Chinese English L2 learners. *Cognition, 87*(2), 129-149.

- Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130-163.
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46-65.
- Webb, S., & Chang, A. C. (2012). Vocabulary learning through assisted and unassisted repeated reading. *Canadian Modern Language Review, 68*(3), 267-290.
- Webb, S., & Chang, A. C. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research, 19*(6), 667-686.
- Webb, S. & Chang, A. (2015). Second language vocabulary learning through extensive reading with audio support: how does frequency and distribution of occurrence affect learning? *Language Teaching Research, 19*(6), 667-686.
- Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.
- Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal, 104*(4), 715-738.
- Wesche, M., & Paribakht, S. (1996). Assessing second language vocabulary knowledge: depth versus breadth. *Canadian Modern Language Review, 53*(1) 13-40.
- Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of experimental psychology: learning, memory, and cognition, 15*(6), 1047.
- Wray, A. (2002). *Formulaic Language and The Lexicon*. Cambridge: Cambridge University Press.

- Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning*, 71(2), 487-536.
- Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: effects of frequency and contextual richness. *The Canadian Modern Language Review* 57(4), 541-572.
- Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System*, 33(4), 547-562.
- Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 26(4), 696-725.
- Ziegler, J. C., & Goswami, U. (2006). Becoming literate in different languages: similar problems, different solutions. *Developmental science*, 9(5), 429-436.
- Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal reading and predictors of reading: a cross-language investigation. *Psychological Science*, 21(4), 551-559. <https://doi.org/10.1177/0956797610363406>