A CONSTRUCT VALIDATION STUDY OF IMPLICIT AND TIME SENSITIVE VOCABULARY MEASURES

By

Bronson Hui

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies—Doctor of Philosophy

2021

ABSTRACT

A CONSTRUCT VALIDATION STUDY OF IMPLICIT AND TIME SENSITIVE VOCABULARY MEASURES

By

Bronson Hui

Vocabulary researchers have started expanding their assessment toolbox by incorporating timed tasks and psycholinguistic instruments (e.g., priming tasks) to gain insights into lexical development (e.g., Elgort, 2011; Godfroid, 2020b; Nakata & Elgort, 2020; Vandenberghe et al., 2021). These timed sensitive and implicit word measures differ qualitatively from traditional paper- or accuracy-based vocabulary tests and are believed to tap into lexical strength and representations in the mental lexicon (Elgort, 2018; Godfroid, 2020b). As a result, there have been calls to use both traditional (explicit) and these timed and implicit word measures in a complementary manner (e.g., Godfroid, 2020b; Nakata & Elgort, 2020; Vandenberghe et al., 2021). At the same time, researchers must first develop a thorough understanding of how these different types of measures (explicit *vs.* implicit and timed *vs.* untimed) relate to each other before they can make informed decisions on their measurement battery. It is thus well-motivated to examine the construct validity of these measures empirically and systematically. In this validation study, I took the first step to fill this research gap by assessing both the predictive and factorial structure validity of these measures.

One hundred and forty-five learners of English took part in five vocabulary tasks: (1) a receptive form-meaning task, where they chose an option representing the meaning of the target word embedded in a sentence, (2) a productive form-meaning task, where they

produced the target word to fit a sentence context, (3) a computerized Yes-No (reaction time) test, where they indicated if they knew the target word by pressing keys on their keyboard, (4) a masked repetition priming task with lexical decisions, where they judged if a letter string forms a word in English, and (5) a semantic priming task with lexical decisions. Items in all five tests were the same 40 English words sampled across the 2K - 5K frequency bands.

Data analysis involved item inspection and extraction of person-related parameters based on Rasch and/or mixed-effects models. The measures of person ability obtained from individual tasks were then submitted to confirmatory factor analyses in order to assess the psychometric dimensionality of the measure battery. The resulting latent factor(s), representing a pure measure of vocabulary under a specific conceptualization, was then used to predict self-reported proficiency to shed light on their predictive validity. With method effects accounted for, the one-factor solution ("Vocabulary Knowledge") produces a good fit and is preferred based on the principle of parsimony for both the implicit *vs*. explicit and timed *vs.* untimed distinctions. This result provides evidence for psychometric unidimensionality of these measures as representing a potential unitary construct of vocabulary knowledge. At the same time, the vocabulary construct has the most explanatory power (predictive validity) when conceptualized distinctly as lexical knowledge (measured untimed tasks) and strength (measured by timed tasks). Taken together, these results foreground the need for researchers to further specify the nature of the vocabulary construct as well as the operational task features with which it can be assessed empirically. Importantly, I call for more measurement validation work as researchers expand their assessment toolboxes in vocabulary research.

To the brave and freedom-loving people of Hong Kong

ACKNOWLEDGEMENTS

I would like to express my deepest possible gratitude to the following individuals for their support and guidance throughout my PhD career.

I am forever indebted to my family for their support and love. This journey would not have been possible without them. I have always been encouraged to adventure and leave my comfort zone. Together, we have done exactly that, sailing away from the place we once called home.

Many thanks go to my advisor, Dr. Aline Godfroid, who has been tremendously supportive. She is certainly a role model and a great mentor, guiding me through the challenges in academia. Her insights often mean more work on my end, but the results are always amazing. I must also thank members of my dissertation committee: Drs. Shawn Loewen, Paula Winke, Patti Spinner, Irina Elgort, and Hope Akaeze. The wide range of expertise they have collectively offered has greatly strengthen the investigation in this work.

Last but not least, I am grateful for the following financial support: the *Language Learning* Dissertation Grant, the Gorilla Grant for Graduate Students, the Dissertation Completion Fellowship, the Graduate School's Research Enhancement Scheme, and the research and conference support from the Second Language Studies (SLS) program.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xi

INTRODUCTION 1

CHAPTER 1: LITERATURE REVIEW 4
Conceptualizations of Vocabulary Knowledge 4
Vocabulary Breadth, Depth, and Strength 4
Implicit and Explicit Word Knowledge 12
Measures of Vocabulary Knowledge 16
Traditional and Time-Sensitive Word Measures 17
Timed Lexical Measure of Strength 21
Implicit Word Measures 23
Measurement Validation 28
Measurement Validation Studies in Grammar Research 30
Measurement Validation Studies in Vocabulary Research 32
The Present Study 36

CHAPTER 2: METHODOLOGY 39
Participants 39
Critical Words 42
Data Collection Platform 44
Measures 49
Form-Meaning Receptive Test 50
Form-Meaning Productive Test 52
Yes-No RT Test (Access to the Form-Meaning Link) 53
Masked Repetition Priming (Lexical Representations) 56
Semantic Priming (Semantic Representations) 63
Self-Reported Proficiency 73
Procedure 74
Data Analysis 75
The Form-Meaning Receptive Test 76
The Form-Meaning Productive Test 79
Yes-No Rt Test 80
Masked Repetition Priming 82
Semantic Priming 86
Main Cfas And Sems 86
Analysis Software Packages 92

CHAPTER 3: RESULTS (INDIVIDUAL TASKS) 93
The Form-Meaning Receptive Test 93
The Form-Meaning Productive Test 95
Yes-No RT Test 101
Accuracy Data 101
Reaction Time Data 103
Masked Repetition Priming 106
Semantic Priming 107
Summary of Results for Individual Tasks 109

CHAPTER 4: RESULTS (CFA AND SEM) 112
RQ1a – Explicit *vs.* Implicit 112
RQ1b – Knowledge *vs.* Strength 112
RQ2a – Predictive Validity of a Single Vocabulary Construct 117
RQ2b – Predictive Validity of Lexical Knowledge and Strength 117
Summary of Findings 118

CHAPTER 5: DISCUSSION AND CONCLUSION 119
The Jury is out but… 119
A Broader, Unitary View of Vocabulary Knowledge 120
What Implicit and Timed Measures Offer 123
Understanding Priming Tasks as Individual Differences Measures 125
Alternative and Equivalent Models 129
Limitations and Future Directions 131
Conclusion 132

APPENDICES 134
APPENDIX A THE FORM-MEANING RECEPTIVE TEST 135
APPENDIX B THE FORM-MEANING PRODUCTIVE TEST 150
APPENDIX C STIMULI FOR THE YES-NO RT TEST 153
APPENDIX D STIMULI FOR THE MASKED REPETITION PRIMING TASK 155
APPENDIX E WORD ASSOCIATION NORMS FOR CRITICAL RELATED TRIALS 170
APPENDIX F STIMULI FOR THE MASKED SEMANTIC PRIMING TASK 173

REFERENCES 186

LIST OF TABLES

Table 1 Nation's (2013) Framework of Vocabulary Knowledge 7

Table 2 Desired Sample Sizes Based on Model Fit..... 40

Table 3 Demographic Information About the Participants 42

Table 4 List of Critical Words 44

Table 5 Summary of Measures for the Present Research 50

Table 6 Means and Standard Deviations of Reaction Times for Words and Non-words 56

Table 7 Summary of Trial Types in the Masked Repetition Priming Task 58

Table 8 Means and Standard Deviations of Reaction Times for Critical Words Between Conditions (Repetition Priming)..... 60

Table 9 Summary of Mixed Models for Pilot Data - Masked Repetition Priming Task 61

Table 10 Summary of Trial Types in the Masked Semantic Priming Task 67

Table 11 *Means and Standard Deviations of Reaction Time for Learners in the Semantic Priming Task - Piloting* 69

Table 12 Means and Standard Deviations of Reaction Time for Native Speakers in the Semantic Priming Task - Piloting 71

Table 13 Summary of Mixed Models for Pilot Data – Semantic Priming Task (Native Speaker). 72

Table 14 Summary of Number of Data Points for Each Participant 75

Table 15 Summary of Hypothesized CFA Models 89

Table 16 Summary of Software Packages Used..... 92

Table 17 Descriptives for the Form-Meaning Receptive Test..... 93

Table 18 Descriptives for the Form-Meaning Productive Test 96

Table 19 Descriptives for the Yes-No RT Test (Accuracy Data)..... 101

Table 20 Descriptive Statistics for the Yes-No Test (Reaction Time Data) 103

Table 21 Descriptive Statistics for the Masked Repetition Priming Task 106

Table 22 Summary of Mixed Models - Masked Repetition Priming Task 107

Table 23 Means and Standard Deviations of Reaction Times for Critical Words Between Conditions (Semantic Priming) 108

Table 24 Summary of Mixed Models - Semantic Priming Task 109

Table 25 Summary of Individual Task Results 110

Table 26 Correlation Matrix for the Individual Task Results 111

Table 27 Summary of Confirmatory Factor Analysis and Structural Equation Models 114

Table 28 Model Summary of CFA-M2 115

Table 29 Model Summary of CFA-M3 116

LIST OF FIGURES

Figure 1 The Modified Hierarchical Model (Pavlenko, 2009) 11

Figure 2 Visualization of CFA-M2 91

Figure 3 Visualization of SEM-M3 91

Figure 4 Person-Item Map for the Receptive Test (40-Item)..... 98

Figure 5 Person-Item Map for the Receptive Test (35-Item)..... 99

Figure 6 Person-Item Map for the Productive Test (38-Item) 100

Figure 7 Person-Item Map for the Yes-No RT Test - Word Data (33-item) 104

Figure 8 Person-Item Map for the Yes-No RT Test - Non-Word Data (37-item) 105

INTRODUCTION

Vocabulary knowledge has been consistently found to be the most important determinant of success in second language use such as reading and listening (e.g., S. Zhang & Zhang, 2020). At the same time, the construct of vocabulary knowledge has been conceptualized in many different ways (e.g., Schmitt, 2014; Yanagisawa & Webb, 2020). These conceptualizations carry different theoretical implications for what it means to know a word. Importantly, each conceptualization requires valid measurement tools for teachers, researchers, and language testers to assess different aspects of learners' word knowledge. Traditionally, vocabulary is often assessed through paper-and-pencil, accuracy-based tests. In a complementary manner, researchers have recently started using reaction-time-based, psycholinguistic tasks in vocabulary studies (e.g., Elgort, 2011; Godfroid, 2020b; Nakata & Elgort, 2020; Vandenberghe et al., 2021). Unlike traditional, explicit vocabulary tests, these implicit and time sensitive word measures are believed to tap into learners' lexical strength and representations in the mental lexicon, respectively. Due to the qualitative differences between these measures their traditional counterparts, they can potentially shed new light on the vocabulary construct and the acquisition process. Therefore, the adoption of these measures and the corresponding expansion of the methodological toolbox in vocabulary research bring exciting opportunities. At the same time, the diversifying of assessment tools also calls for a thorough understanding of the relationships between the many different measures. Specifically, researchers need to assess the alignment between their conceptualization of the vocabulary construct and the measurement tools they deploy. Simply put, the question is whether tests used to measure vocabulary knowledge adequately represent the construct of

vocabulary knowledge as it is conceptualized. If not, researchers need to find new measurement operationalizations to tap into the specific construct(s) in question. Alternatively, the construct might require a reconceptualization based on what can be empirically measured. In addition, researchers should quantify and assess the value of the new insights brought about by these implicit and time sensitive word measures. Essentially, what can researchers gain from administering these tests on top of traditional vocabulary tasks? For example, to what extent can researchers better explain the individual differences in language performance? How important is the knowledge measured by these tasks in authentic language use, after considering what is tapped into by traditional measures? Taken together, a construct validation study of vocabulary measurement, which examines psychometric dimensionality and predictive validity, represents an important step forward, offering initial insights into these test-knowledge relationships and provide foundational support for different forms of vocabulary assessment.

The present dissertation is such an attempt at construct validation. It is organized in five chapters. Following this introduction, I first provide a narrative review of the literature in Chapter 1, where I highlight the research gaps that motivated the study and present the research question that guided the investigation. In Chapter 2, I detail the methodology used to address the research question, including my participants, critical words, measures, and data analysis procedure. In Chapters 3 and 4, I report the results for the individual tasks and for the overall confirmatory factor analyses (CFA) and structural equation models (SEM), respectively. In the closing Chapter 5, I discuss the findings in terms of methodology and what they mean for the understanding of the vocabulary construct, before drawing a conclusion. In doing so, I hope

to draw vocabulary researchers' attention to the validity of the measures that they rely on in their studies and call for more validation research in this research area.

CHAPTER 1: LITERATURE REVIEW

Throughout this chapter, I provide a narrative review of the literature on the conceptualizations of word knowledge, vocabulary measures used by second language (L2) researchers, and measurement validation studies in L2 research. The goal of this chapter is to offer a concise, state-of-the-art overview of the research field, through which I highlight the motivation for the present study. First, I will start with the theoretical conceptualizations of vocabulary knowledge, covering vocabulary size, depth, and strength as well as the distinction between explicit and implicit word knowledge. I will highlight some competing conceptualizations and the practical and theoretical needs to empirically test and differentiate between them. Second, I will then discuss implicit and time sensitive vocabulary measures their application in research as well as their limitations. I will also present three example tests which researchers have used: one of timed vocabulary tasks and two implicit word measures. I stress that, despite the variety of available measures, researchers need to engage in measurement validation work to clarify how these measures may or may not relate to each other and to provide validity evidence for their proper use and interpretations. In the third section, I review the literature of measurement validation studies in grammar research followed by the two validation studies published thus far in the domain of vocabulary research. At the end of the chapter, I will present my research question for the present dissertation.

Conceptualizations of Vocabulary Knowledge

Vocabulary Breadth, Depth, and Strength

Breadth and depth is one distinction vocabulary researchers make in conceptualizing lexical knowledge (e.g., Anderson & Freebody, 1981; Schmitt, 2014; Yanagisawa & Webb,

2020). Vocabulary *breadth* refers to the size of one's vocabulary: it is the number of words for which an individual knows "at least some of the significant aspects of meaning" (Anderson & Freebody, 1981, p. 92). Despite the focus on quantity, the definition inevitably required some specification of the quality of the knowledge (i.e., what it means to know a word). In other words, breadth cannot be easily defined independent of depth. This second, quality dimension of word knowledge has been referred to as the *depth* of understanding of a word. This distinction between vocabulary breadth (quantity or size) and depth (quality) has allowed researchers to address research questions on, for example, how many words one needs to comprehend a text (e.g., Laufer, 1992), and on the effects of various vocabulary activities on different aspects of word knowledge (e.g., Webb, 2007), as well as the relationship between size and depth (e.g., Schmitt, 2014).

In terms of the conceptualizations of vocabulary depth, Yanagisawa and Webb (2020) pointed out that there have been many ways in which depth is defined. Some researchers used depth and quality of knowledge in an interchangeable manner (Anderson & Freebody, 1981; Read, 1993). Offering more specificity to the notion of word knowledge quality, Wesche and Paribakht (1996) differentiate "kinds of knowledge of specific words" and "degrees of such knowledge" (p. 13). In other words, there are various sub-components of word knowledge (e.g., meaning and form). In addition, there is a notion of mastery of such knowledge (e.g., how well one understands the meaning of a word). These two different, but related, conceptualizations of depth represent the key approaches researchers take in examining vocabulary depth.

First, conceptualizing depth as component knowledge, researchers break down word knowledge into its different component elements (Henriksen, 1999; Read, 2000; Schmitt, 2014;

Yanagisawa & Webb, 2020). To date, Nation's (2013) framework is the most comprehensive and oft-cited listing of the various components involved in knowing a word (see Table 1). For example, knowing a word means knowing its form, meaning, and use, which are broken down into different aspects such as spoken and written word forms. Each of these aspects is further broken down into receptive and productive knowledge. For example, having receptive knowledge of a word's spoken form is to know (recognize) how the word sounds. More generally, then, the greatest vocabulary depth one can attain, within this framework, is the mastery of all these 18 different components. Due to the comprehensiveness of this framework, it has informed the work of many vocabulary researchers investigating vocabulary depth. For example, in Schmitt's (2014) review of the conceptualizations of the vocabulary depth in research, six out of seven operationalizations can be mapped more or less directly to a component in Nation's (2013) framework. They include receptive versus productive mastery, knowledge of multiple word knowledge components, knowledge of polysemous meaning senses, knowledge of derivative forms (word family members), knowledge of collocation, and the degree, and kind of lexical organization (Schmitt, 2014). Despite the huge popularity of Nation's (2013) framework, one limitation of taking depth exclusively as word component knowledge is that mastery of knowledge is often seen as a binary. For example, does a learner know the spoken form of the word, yes or no? In this light, word component knowledge cannot be easily mapped onto mastery of knowledge, which should be viewed as a continuum (i.e., how well a learner knows [a component of] the word), especially for researchers interested in examining the developmental trajectory of vocabulary knowledge.

Table 1

Nation's (2013) Framework of Vocabulary Knowledge

Aspects of vocabulary knowledge		Receptive/ Productive	What it means to master this aspect?
form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express the meaning?
meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	association	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

Another approach to conceptualizing depth is to view it from a developmental perspective, "from no knowledge to fully developed knowledge" (Yanagisawa & Webb, 2020, p. 373). In other words, a learner can be described in terms of how they develop knowledge from partial to precise knowledge of a word's meaning (Henriksen, 1999; Read, 2004). For example, this progress can mean knowing a word (e.g., *pretty*) is a positive adjective, before knowing its shared meaning with a close synonym (e.g., *beautiful*). Eventually, the learner understands some subtle differences between the two (e.g., *pretty* focuses more on the attractive appearance of a person while *beautiful* also implies a person's positive inner quality). Similarly, lexical development can also be operationalized as progressing from receptive to productive knowledge, whereby a learner may first recognize a word and know its meaning before they can use it appropriately in terms of grammar and meaning (e.g., Henriksen, 1999; Paribakht & Wesche, 1993; Wesche & Paribakht, 1996). While this receptive-productive trajectory has been a common operationalization of development (e.g., Yanagisawa & Webb, 2020), this continuum implies that the end point of development be an accurate and grammatical production of the word.

Yet one may wonder whether one's lexical development truly stops at word production. If not, what should be the target for learners? One suggestion is that increasing vocabulary fluency (or automaticity) "should be the ultimate goal for most language learners" (Qian & Lin, 2020, p. 68). In Schmitt's (2014) words, "[a] way of thinking about what learners can do with lexical items is how fluently and automatically the items can be used in each of the four skills (reading, writing, listening, and speaking)" (p. 920). Fluent and automatic use requires what researchers refer to as strength of lexical knowledge (Nation & Webb, 2011; Webb, 2012;

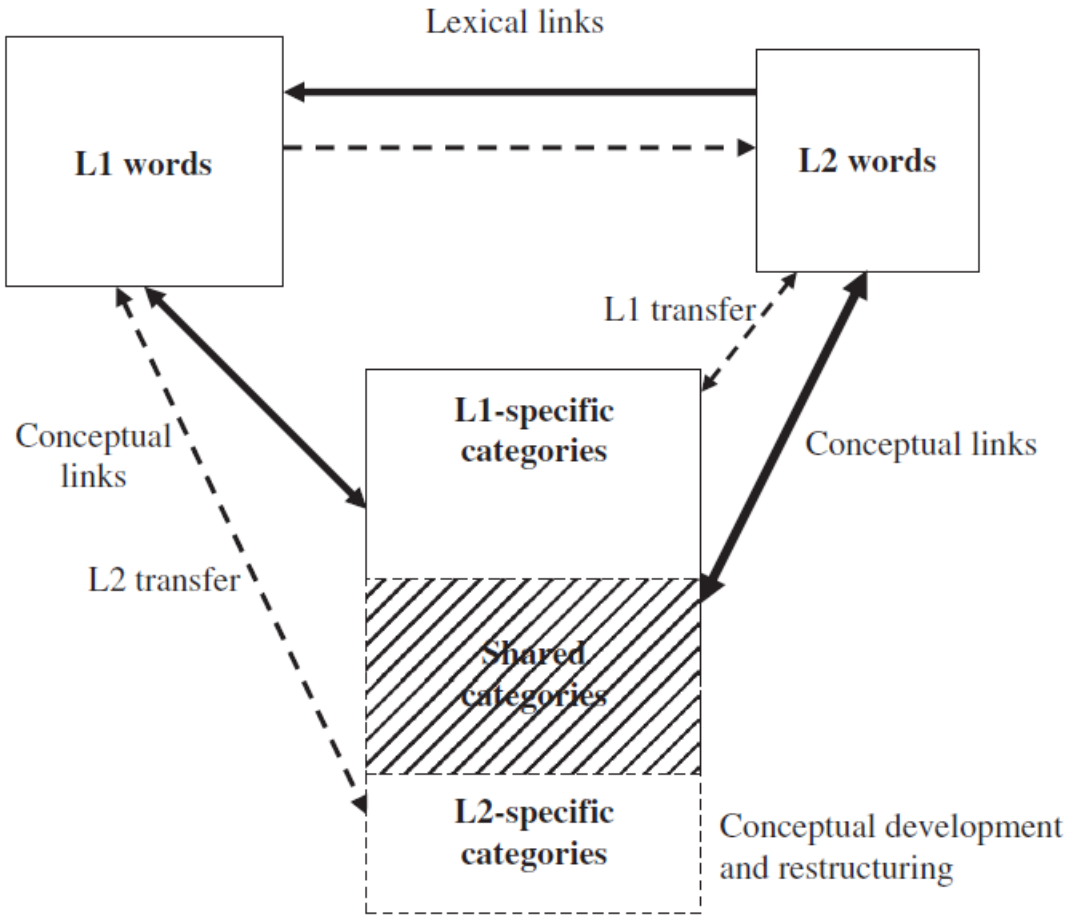
Yanagisawa & Webb, 2020). Strong lexical knowledge is a prerequisite for efficient retrieval and access (Godfroid, 2020b; Yanagisawa & Webb, 2020). Similarly, Perfetti (2007) suggested that adding effective practice to lexical knowledge results in "[processing] efficiency: the rapid, low-resources retrieval" (p. 359). In this light, then, lexical strength represents as an important element of the quality of word knowledge, and it should be distinguished from component word knowledge described earlier (Yanagisawa & Webb, 2020). Again, it is a difference between how well one can master a specific aspect of a word (e.g., form-meaning mapping) and how many different aspects of a word one knows (e.g., meaning, collocation). This isolated treatment of lexical strength and fluent use from word component knowledge echoes Daller et al. (2007) who proposed a three-dimension lexical space with breadth, depth, and fluency. Harrington's (2018) notion of lexical facility similarly encapsulates accuracy, speed, and consistency of access and retrieval, all of which are believed to be the cognitive bases of fluent language use (e.g., Segalowitz & Segalowitz, 1993). Perfetti (2007) also explicitly spelled out the processing consequences of high lexical quality, such as processing stability and synchronicity. Most recently, Godfroid (2020b) proposed to formalize the dimension of fluency of use by expanding Nation's (2013) framework of word knowledge with an additional facet of automaticity. Together, these authors rightly pointed out the importance of considering how well one can use their word knowledge in terms of ease of processing.

Note, however, that access and retrieval are two language processing operations. As rightly pointed out by Perfetti (2007), efficient access is a processing *consequence* of strong, high quality lexical knowledge. It is well documented in the psycholinguistics literature that language processing (e.g., word recognition) can be influenced many factors other than the

quality of lexical representations in the mind. These factors include one's language dominance, differences and similarities between the two languages of a bilingual, frequency of use and so on (e.g., Kroll & Tokowicz, 2005). Indeed, psycholinguistic models of the bilingual lexicon often distinguish knowledge representations in memory from processing operations. For example, in the Modified Hierarchical Model (Pavlenko, 2009), second language learning is seen as developing and restructuring conceptual categories that can be L1-specific, L2-specific, or shared (see Figure 1). According to this model, one learning task for L2 speakers is then to disambiguate subtle differences at the conceptual level between and within the two languages. According to Pavlenko (2009) this learning is "a gradual process, taking place in implicit memory" (p. 159), which I will return to in the next section. The key here is that conceptual knowledge, together with the lexical, formal knowledge, are represented in memory. Strong, robust representations are easier to access and retrieve during recognition and production. At the same time, these language processing operations are often carried out under the mutual interaction between the mind and the environment. For example, in most word recognition models, when a word is recognized, its representation needs to have a high level of activation that surpasses that of other word candidate (e.g., Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986). This activation level during the recognition process is shaped by, for example, both the extent to which the representations are precise (the mind) and whether the language is being used exclusively in the context (the environment). The bottom line here is, in addition to strong lexical knowledge, which vocabulary researchers are most interested in, processing-related factors influencing activation levels during online processing also play important roles in how vocabulary is used in communication.

Figure 1

The Modified Hierarchical Model (Pavlenko, 2009)



To summarize this section, I have reviewed the contrast between vocabulary size (focusing on knowledge quantity) and depth (emphasizing knowledge quality). Depth in turn should be conceptualized in two distinct ways: first, knowing the different components of a word (e.g., form, meaning, and use); second, developing strength of such knowledge. It is this second way, of strength, that is the fundamental basis of real-life, efficient access and retrieval in authentic communication settings although other factors can also influence processing. On

this account, teachers, researchers, and language testers should focus more on the teaching, learning, and measuring of vocabulary strength. At the same time, there are theoretical questions that need to be addressed. For example, what is the exact relationship between word knowledge and lexical strength? To what extent is vocabulary knowledge as strength a separate dimension from word knowledge? If they are separate, how do they develop? Clarifying these questions will help researchers develop a unified theory of word knowledge development in L2 speakers, which in turn will have pedagogical implications. In particular, a solid theoretical understanding will aid teachers when deciding when to conduct classroom activities to promote fluency. For example, if the development of strength follows a separate trajectory, a teacher may wish to start fluency training early to allow different types of knowledge to develop simultaneously. If strength is an extension of the receptive-productive knowledge continuum, a teacher may delay fluency training until students have some solid receptive and productive knowledge first.

Implicit and Explicit Word Knowledge

In addition to the distinction between lexical knowledge and strength, researchers have also suggested that there are "two types of [lexical] knowledge" (Nakata & Elgort, 2020, p. 6). These two types have been labelled as explicit *vs.* implicit, declarative *vs.* procedural (or non-declarative), and available online *vs.* offline (e.g., Nakata & Elgort, 2020). This distinction echoes a similar contrast in grammar research, where researchers distinguish explicit and implicit knowledge (e.g., Andringa & Rebuschat, 2015; DeKeyser, 2003; N. Ellis, 2005; Hulstijn, 2005). In a seminal paper, R. Ellis (2005) summarized the characterizations of these knowledge types and proposed a number of operational features in tasks that could be used to measure them. One

important feature is particularly relevant to the present context: learners' awareness of the knowledge. In particular, the learner always has awareness of their explicit knowledge ("they know that they know"), but not their implicit knowledge.

At the same time, it is crucial to acknowledge that this criterion used to differentiate explicit and implicit grammar knowledge may not be directly applicable to vocabulary research (Sonbul & Schmitt, 2013). As far as the form-meaning link is concerned, for instance, it may be difficult to conceive that a learner has no awareness of a word's meaning. Indeed, the arbitrary pairings between form and meaning are believed to reside in declarative memory along with other information that the learner is conscious of (e.g., Ullman, 2001). Hulstijn (2007) similarly suggested that vocabulary knowledge is largely explicit because of its symbolic nature (i.e., the form-meaning relationship is arbitrary). For both N. Ellis (1994) and R. Ellis (2004), the semantic components of a word's knowledge are often explicit, while knowledge related to form and use can be either explicit or implicit. Similarly, Sonbul and Schmitt (2013) pointed out the word's relationship to the broader linguistic system of the knowledge should be factored into the consideration of explicit and implicit vocabulary knowledge. In sum, the semantics associated with vocabulary knowledge has led researchers to suggest that word knowledge is largely explicit in nature.

While there may be a lot of truth in the proposition, the picture is far more complex. First, as briefly mentioned, psycholinguistic models of the bilingual lexicon have attempted to incorporate lexical items and concepts in *both* languages (e.g., Pavlenko, 2009). While a learner can often verbalize a translation equivalent of a word, they might not be aware of the subtly restructured (L2-specific) conceptual representations in memory that results from extensive

language use. Indeed, as acknowledged by Pavlenko (2009), the distinction between implicit and explicit word knowledge "has not yet been incorporated into models of the bilingual lexicon" (p. 150). Therefore, concluding that the semantic components of word knowledge is categorically explicit may be oversimplistic. In addition, the view of lexical knowledge being explicit stems from a narrow definition of what it means to know a word (i.e., knowing a word is only about knowing its meaning). Both SLA researchers (e.g., Meara, 2009; Meara, 1992; Nation, 2013) and psycholinguists (e.g., Jiang, 2015; McNamara, 2005) have viewed the semantic components of lexical knowledge as an interconnected network. For example, in Nation's (2013) framework, mastering meaning association at the productive level means knowing what other words can be used in a given context to convey a similar meaning. Meara (1992, 2009) also views the lexicon as a network structure where items are connected through associations between them. In psycholinguistics, it is found that activation of a lexical item can spread to other items via a connected, semantic network (e.g., McNamara, 2005). On this account, researchers need to differentiate how well a learner knows a particular item (e.g., its meaning) from how well the item in question is connected to other items (or how well it is integrated in the lexicon). It appears that researchers suggesting that vocabulary knowledge is explicit place more emphasis on the former. However, the latter view of knowing a word (i.e., in terms of integration into the lexicon) is perhaps more comprehensive and deserves more attention in vocabulary research.

With regard to learners' awareness of such integration in the lexicon, the issue is also complicated. On the one hand, knowledge of meaning association can be inside a learner's awareness. For example, a learner can tell that *nurse* and *doctor* are related semantically. On

the other, psycholinguistic research has also shown that certain knowledge can be involved in language processes that are "not available to learners' conscious control or report" (Elgort, 2018, p. 4). In other words, when a task does not invoke the learner's awareness of the knowledge, what is being measured might be considered as implicit. For example, in a psycholinguistic experiment, a brief presentation of a word (e.g., *nurse*) can improve the recognition and/or production of a semantically related word (e.g., *doctor*) (e.g., Collins & Loftus, 1975; McRae & Boisvert, 1998). Since the presentation of the first word (i.e., the prime) is very brief, learners often report a lack of awareness of its presentation. In this light, the processing of *nurse* can be outside the awareness of the learner, but it has implications on the recognition of a subsequent item (*doctor*). Such a task has been used to investigate the extent to which the mental lexicon is interconnected, and more importantly, no awareness of the meaning association needs to be invoked for the effects to be observed. When that is the case, at least some aspects of this meaning association between *nurse* and *doctor* may be implicit. Implicit knowledge such as meaning association between word items can be important in actual language use. According to Nakata and Elgort (2020), for example, implicit or tacit word knowledge "is… needed in fluent, low-effort access to contextually relevant meanings during reading" (p. 7). In the context of listening, a well-established semantic network might help listeners predict upcoming information (e.g., Altmann & Kamide, 1999), potentially lessening some processing burden that can cause a breakdown. However, using awareness as the only criterion to differentiate explicit from implicit knowledge can be "problematic and thorny" even in grammar research (Leow, 2001, p. 118). Recent evidence has also shown that different operationalizations of (un)awareness can become unreliable, leading to potentially different

conclusions (Maie & DeKeyser, 2020). Therefore, caution should be exercised when defining explicit *vs.* implicit knowledge in terms of learner awareness.

In this section, I reviewed two key distinctions in conceptualizing the vocabulary construct. The first relates to the contrast between lexical knowledge and strength. The former concerns what a learner knows, and the latter represents the degree of mastery. I also discussed the distinction between explicit and implicit word knowledge and the use of (un)awareness as a criterion to distinguish the two. In the current literature, what might be less clear is how lexical strength is related to implicit knowledge, or if they are related at all. However, one potential link between the two might be ease of access. As established above, ease of access is a key characterization of lexical strength (e.g., Daller et al., 2007; Harrington, 2018). Similarly, implicit knowledge can be accessed relatively effortlessly (e.g., R. Ellis, 2005; Nakata & Elgort, 2020). In this light, then, ease of access can serve as a common ground for lexical strength and implicit word knowledge, representing one criterion in the distinction between (explicit) lexical knowledge and word knowledge strength, the latter of which encapsulates implicit knowledge. Given these theoretical conceptualizations and hypothesized dimensionality, I review the measurement of these dimensions of the vocabulary construct below.

Measures of Vocabulary Knowledge

In this brief overview of vocabulary measures, I will first start with discussing two types of "qualitatively different" measures: traditional and time-sensitive measures (Godfroid, 2020b, p. 433). Then, I will discuss timed measure of lexical strength and two implicit word measures,

which can potentially help researchers empirically operationalize constructs such as vocabulary knowledge strength and implicit word knowledge discussed in the previous section.

Traditional and Time-Sensitive Word Measures

There have been a number of related terms that researchers have used to contrast with traditional, paper-based vocabulary tests. These terms include online *vs.* offline, implicit *vs.* explicit, time-sensitive *vs.* paper-based, but they refer to different sets of tests despite some overlaps. Although the use of these terms are not always consistent in the literature, three criteria can be used to distinguish them: (1) is it a real-time measure?, (2) is there time pressure imposed on the participant?, and (3) does the task invoke awareness of the knowledge?

First, Godfroid (2020b) defines online measures as those that tap into "learners' lexical knowledge *during* language processing (hence the name online) when there are real time restrictions" (p. 433, emphasis added). The author compared online measures to watching sports in real time where "the detailed happenings of the match… unfold" (p. 433). In other words, these tests are characterized by a real-time element and a time-restriction component. Eye tracking is an example of real-time measures in the sense that eye movements are recorded as the participant is engaged in language use such as reading and listening (see (Godfroid, 2020a for an overview). With regard to time pressure, timed tasks as in a psycholinguistic experiment, which collects reaction time data, require participants to respond quickly on a trial-by-trial basis. Time pressure in psycholinguistic is commonly operationalized in the task instructions which ask participants to respond as fast as possible and/or in the programming of the experiment where a given trial will end after a set amount of time. Also, this pressure is intentionally placed on the participant, so they do not have time to reflect on

their responses. As a result, researchers can infer the participant's access to knowledge and potentially the nature of knowledge. On this account, this time pressure serves an explicit goal in the experimental operationalization and hence is different from a time limit in typical language testing contexts. Often test administers would, on the one hand, standardize engagement by imposing a time limit; and on the other, ensure the vast majority of test takers have the time to complete the assessment without inducing test-irrelevant variance (e.g., that results from test anxiety) (e.g., Denovan & Dagnall, 2019). Finally, implicit measures target knowledge that is involved in language processes that are outside the learner's awareness. As mentioned, it might involve a very brief presentation of linguistic materials such that the participant is not aware of the presentation. A broader term that encapsulates these tests is sensitive measures. In the present dissertation, I use sensitive measures as an umbrella term to refer to vocabulary tests that impose time pressure on the test taker and those that target implicit word knowledge. These tests are contrasted with traditional vocabulary tests that are explicit, offline, and untimed.

As already alluded to, these task features are in sharp contrast with traditional, paper-based, untimed, explicit word measures that have been used predominated in the field of vocabulary acquisition (Godfroid, 2020b). These traditional tests are of different formats ranging from multiple-choice questions, translation (first to second or second to first language) tasks, to fill-in-the-blank and matching items. They can be designed so as to target different word knowledge components in Nation's (2013) framework. For example, in the 14k Vocabulary Size Test (Nation & Beglar, 2007), participants are presented a target lexical item and a neutral sentence context from which the meaning cannot be inferred. The task for the

participant is to choose the correct, corresponding meaning out of four options (see an example item in the Measures section in Chapter 2). Another example is the productive Vocabulary Levels Test (Laufer & Nation, 1999) where participants are presented with a sentence context for a missing target word. The task for the test taker is to fill in the blank with a word that fits the context (see an example item in the Measures section in Chapter 2).

These offline tests have made a significant contribution to vocabulary research. They have allowed researchers to identify, for example, the relationship between language use and vocabulary (e.g., Cheng & Matthews, 2018; Jeon & Yamashita, 2014; Vandergrift & Baker, 2015). For example, Vandergrift and Baker (2015) showed that second language (L2) vocabulary was most directly and strongly associated with L2 listening performance. Similarly, in reading research, Jeon and Yamashita (2014) meta analyzed 31 independent correlations between L2 reading comprehension and L2 vocabulary knowledge from 29 studies. The authors found a high correlation of .79 with 95%CI[.69 – .86]. In addition, these offline tests also provide an outcome measure for researchers to understand how different learning conditions (e.g., various types of glossing) can impact vocabulary learning (e.g., H. S. Kim et al., 2020; Ramezanali et al., 2021; Yanagisawa et al., 2020). For example, Yanagisawa et al. (2020) meta-analyzed 359 effect sizes in 42 studies and found an overall advantage of glossing (*vs.* no glossing), with multiple choice glosses (where only one of the different senses presented fits the current context) being the most effective.

Despite the contribution of these tests, Godfroid (2020b) argued that researchers should consider the face validity of these tests. For example, one should assess the alignment between how vocabulary knowledge is measured and how it is used in real-life communication.

L2 listening comprehension is a case in point. Since listeners have relatively less control over the speed of the incoming speech stream (Hui & Godfroid, 2020; K. M. Kim & Godfroid, 2019), the efficiency with which the listener's cognitive processor can manage the flood of information can be a key to successful comprehension (Hui & Godfroid, 2020; Vafaee & Suzuki, 2020). On this account, possessing word knowledge of spoken word form and the form-meaning connection, as can be demonstrated on an offline vocabulary test, is perhaps only one necessary condition for comprehension. Efficiently putting that knowledge to use (i.e., efficiently processing phonological and semantic information) may also be necessary. If that is the case, when tested without time pressure, learner's ability to point out the meaning of a word (e.g., in the 14k Vocabulary Size Test) may not entirely coincide with their efficient access to such knowledge in authentic communication. Indeed, recent evidence has shown that performance in a timed lexical task can account for some unique variance in L2 listening comprehension at both propositional and discourse levels (Hui & Godfroid, 2020), meaning that efficient access to lexical knowledge carries explanatory power above and beyond vocabulary size measured by offline tests. Similarly, in reading research, Tanabe (2016) reported that reaction times in a computerized vocabulary test, but not vocabulary test scores (i.e., accuracy scores) alone, were a significant predictor of reading comprehension under time pressured conditions. His results echo the idea that fluent reading requires efficient access to the meaning of words with minimal effort, which can in turn free up cognitive resources for high-level processing (e.g., Elgort et al., 2018; Nakata & Elgort, 2020; Perfetti, 2007). On this account, then, time-sensitive measures that are of a different nature appear to have the potential to carry complementary roles in measuring vocabulary in a comprehensive manner. On the

contrary, using exclusively offline word measures can cause some aspects of lexical skills being under-represented and hence introduces bias to the vocabulary construct (e.g., Révész & Brunfaut, 2021). Importantly, these measures, due to their own unique characterization, provide potential empirical operationalizations of such dimensions as lexical strength and implicit word in the vocabulary construct. Below, I present a brief overview of one timed lexical measure of strength and two implicit lexical tasks.

Timed Lexical Measure of Strength

Yes-No RT Test. Although the Yes-No test format, which requires test takers to select the words they know from a list, dates back to 1929 (Beeckmans et al., 2001), Meara and Buxton (1987) were the first to use this format as a vocabulary test for second language learners. Traditionally, accuracy data to real words and non-words provide measures of vocabulary knowledge and the level of guessing, respectively. In particular, correct responses to real words (hits) are used to infer one's vocabulary knowledge, while the incorrect yes responses to non-words (false alarms) provide information regarding the test taker's level of guessing. A more reliable estimate of one's lexical knowledge often involves both hit and false alarm measures whereby adjustment for guessing is factored into the scoring (see Huibregtse et al., 2002 for a comprehensive review).

This format of two-option forced choices (i.e., Yes *vs.* No) is flexible because it can be seen as aligning with how reaction times are typically measured in psycholinguistic research. For example, in psycholinguistic research, participants can be asked to indicate whether a sentence presented to them is grammatically acceptable or not by pressing the corresponding Yes or No button on a response pad as quickly and accurately as possible (for a methodological

review, see Plonsky et al., 2020). In a lexical decision task, participants decide whether or not a presented letter string forms a word, which is similar to a Yes-No test when programmed to collect reaction times. Indeed, Harrington (2006) proposed using reaction time data from lexical decisions as vocabulary measures to index both accuracy and speed of access to word knowledge. In the study, the author reported that higher proficiency learners tended to respond faster and more accurately. Similarly, Pellicer-Sánchez and Schmitt (2012) collected and analyzed reaction time data on a Yes-No RT test. The authors compared a reaction time-based scoring approach and a non-word approach, whereby they adjusted for guessing based on responses to non-words. For the reaction time-based approach, the authors established reaction time thresholds (e.g., 577.27 ms for non-native speakers) to discriminate between accurate and inaccurate responses. However, the authors found no clear advantage of using the reaction time approach. In Hui and Godfroid (2020), the authors used an auditory Yes-No RT test to investigate the relationship between lexical strength and second language listening comprehension. They used accuracy, reaction time, and the coefficient of variation (CV where $CV_{RT} = \frac{M_{RT}}{SD_{RT}}$) (Segalowitz & Segalowitz, 1993) to index vocabulary size and lexical processing speed and automaticity. Regression and subsequent mediation analyses showed that accuracy and reaction times on the test predicted L2 listening comprehension at both propositional and discourse levels.

To date, use of the Yes-No RT test is still rather limited in SLA research, such a timed test allows researchers to make inferences about one's efficiency of accessing lexical information. At the same time, it is not entirely clear how reaction time data, which are believed to unveil one's lexical strength, are related to other traditional, paper-based tests which entirely ignore

time pressure as a condition in real-life communication. Pellicer-Sánchez and Schmitt (2012) was the first to try to establish such relationships between timed and untimed tests, but more validation work needs to be conducted to scrutinize the underlying construct(s) that the different measures afforded by the test are tapping into.

Implicit Word Measures

Masked Repetition Priming. The priming paradigm is commonly used in psycholinguistic and bilingualism research (e.g., Trofimovich & McDonough, 2011; VanPatten & Jegerski, 2014). In general terms, the mechanism of the paradigm involves prior exposure to linguistic information (i.e., the prime) influencing (often facilitating) subsequent recognition and/or production of language. In the case of masked repetition priming, the facilitation has been a well-establish phenomenon in both L1 and L2 speakers (e.g., Evett & Humphreys, 1981; Forster & Davis, 1984; Gollan et al., 1997; Jiang, 1999). At the behavioral level, participants have been found to respond faster to the target word when it is preceded by an identical prime than when the target word is preceded by an unrelated, non-identical prime. Two aspects of this phenomenon are important: first, the prime is presented only very briefly and is masked (e.g., preceded and/or followed by symbols such as a string of hash signs [#]), often resulting in a reduced visibility of the prime and a lack of awareness by the participant that it is there. Therefore, the processes underlying this phenomenon are considered automatic and operate largely out of the conscious control of the participant. Although the decisions the participant makes about the targets are conscious, the processes that this priming task taps into are not. Second, masked repetition priming is not found on non-word trials, indicating that the processes which drive this effect operate at the lexical level (e.g., Forster, 1998), as opposed to the sub-lexical level

(i.e., lexical forms, or spelling). To account for this effect, then, researchers proposed that, in an experimental trial, the prime pre-activates the lexical representation carrying both formal-lexical and semantic information. This pre-activation in turns facilitates the recognition of the target (e.g., Grainger et al., 2003), leading to faster responses. In the context of vocabulary learning, when a learner has acquired a word in the sense of having established a lexical representation, such facilitation (i.e., priming) should be observed. In contrast, if the lexical representation is not robust, or if no lexical representations have been established, such priming may not be observed. Leveraging this phenomenon, researchers can then examine vocabulary knowledge and acquisition via lexical processes that are not in the learner's awareness. Importantly, this fine-grained measure represents a tool to investigate the extent to which a certain learning condition may shape vocabulary acquisition. For example, when an experimental learning condition is less conducive to learning, a reduced (or eliminated) priming can be expected (e.g., Elgort, 2017).

While masked priming has been well documented in the psycholinguistic literature, Elgort (2011) was the first researcher who used the task as a vocabulary measure in second language research. In the study, the author examined the extent to which deliberate, decontextualized word learning can lead to the development of tacit word knowledge. Participants learned 48 pseudowords using flash cards and word lists. After one week of study, they returned to the lab and took part in three priming tasks: a form priming, a masked repetition priming, and a semantic priming task. Focusing on the masked repetition priming task, results showed that participants responded on average 52 ms and 75 ms faster on the related, repetition trials (than on the unrelated trials) for the newly learned pseudowords and

low-frequency real words, respectively. As expected, no facilitation was found for the non-word trials. The robust priming observed for the newly learned pseudowords led the author to conclude that deliberate, decontextualized word learning can result in tacit word knowledge, measured by the priming task (see also Elgort & Piasecki, 2014).

In another study, Elgort (2017) investigated the extent to which the accuracy of initial word meaning inferences under contextual learning conditions has an impact on subsequent word learning outcomes. Participants were first exposed to target words embedded in informative sentences. They were asked to make inferences based on the context, after which they were presented with the correct meaning for verification. During the testing phase, the participants took part in, among other tests, a mixed-modality priming task where the masked prime was presented visually, while the target was presented auditorily. Results showed robust priming in the sense that participants responded faster on the related trials (where the visual and auditory stimuli corresponded) than on the unrelated trials. But, this priming was not moderated by the accuracy of the initial meaning inferences the learner had made in the learning phase when reading the informative sentences. Based on the lack of a significant interaction, the author concluded that incorrect meaning inferences "appear to be benign as far as the development of implicit knowledge and establishment of lexical representations are concerned" (Elgort, 2017, p. 8).

In sum, the masked repetition priming paradigm has been adopted to L2 vocabulary research to gain insights into the extent to which learners have established a robust lexical entry for the word items tested. An observed priming effect is taken as evidence of implicit

word knowledge and reflects that a lexical representation has been established in the mental lexicon.

Semantic Priming. A semantic priming task is another test in the general priming paradigm commonly used in psycholinguistic research. Semantic priming can be used to evaluate the robustness of the semantic representations of the stimuli. As with the repetition priming, a facilitation (priming, or faster responses) is expected when the prime and the target are semantically related. The idea is that, on related trials, the prime activates the semantic network which overlaps with that of the target. This activation then facilitates the recognition of the target, leading to faster responses. On unrelated trials, since there are no semantic overlaps, no priming should be observed. This semantic priming effect is, again, well documented in psycholinguistic research and is taken as evidence that the semantic information between word items is interlinked (e.g., Collins & Loftus, 1975; McRae & Boisvert, 1998). In vocabulary research, then, priming should be observed when the participant (1) has integrated the stimuli in semantic network, (2) can access these lexical semantic representations fluently, and (3) processes the primes and targets on the related trials as semantically similar (e.g., Elgort & Warren, 2014).

In Elgort's (2011) study, reviewed previously, the author found a 22-ms semantic priming effect, indicating that the participants had acquired the lexical-semantic representations for the pseudowords that they had learned in a deliberate, decontextualized manner. This level of priming was compared with the 37-ms facilitation when the prime was a known English word. The author suggested that "these representations [for the pseudowords] were probably less stable than those of known L2 words and that their integration into the

lexical-semantic memory system of the participants was in its early stages." (Elgort, 2011, p. 394). In another study, learners did not show reliable semantic priming after encountering pseudowords embedded in a text multiple times under incidental word learning conditions (Elgort & Warren, 2014). This result was despite an interaction whereby the degree of priming depended on the age when the participant had started learning the target language (i.e., English). Specifically, there was "faster and more robust lexicalization… for those who started learning English earlier in life" (Elgort & Warren, 2014, p. 396). Similarly, Bordag et al. (2015) and Chen (2021), in two incidental word learning studies, reported a lack of robust semantic representations after multiple exposures to target novel words. Taken together, perhaps, semantic integration represents a high bar for learners as far as word learning is concerned. When learners study words intentionally, semantic priming can be observed. In contrast, if vocabulary learning takes place under incidental conditions, the integration (if any at all) may not be robust enough to be detected by a semantic priming task. In more general, methodological terms, the semantic priming task has been used as another implicit vocabulary measure to examine the extent to which semantic representations of word items have been established in the mental lexicon.

Taken together, the adoption of these time-sensitive measures to the investigation of lexical strength and implicit word knowledge and learning has expanded the assessment toolbox of vocabulary researchers. Researchers are now in a better position to address questions pertaining to implicit word knowledge and efficient access to word knowledge. Importantly, they can answer the calls from vocabulary scholars to use multiple measurements concurrently to better understand the construct of vocabulary knowledge at a theoretical level

(e.g., Milton & Fitzpatrick, 2014; Read, 2020; Schmitt, 2010; Webb, 2005; Yanagisawa & Webb, 2020). At the same time, it is crucial to understand what is being measured exactly by each of these time-sensitive measures. To date, the use of these measures has been advocated through argumentation (e.g., time pressure is present in authentic communications, and so timed tasks need to be administered to add face validity to the measures). Not much work has provided empirical evidence that imposing time pressure qualitatively changes the knowledge being measured and/or that using an implicit task can measure what explicit tests cannot tap into. Perhaps more importantly, researchers must understand how exactly these time-sensitive measures can complement traditional ones. For example, might it be the case that these time-sensitive tests simply represent a more difficult set of tasks (or a higher learning target), but they ultimately tap into the same dimension of word knowledge as that measured by explicit tests? Alternatively, do these reaction time-based tasks represent measures that are qualitatively different from explicit tests? At a practical level, clarifying how the different tools at researchers' disposal relate to each other will enable more informed decisions regarding what sets of measures to administer. In theoretical terms, engaging in measurement validation work makes conceptual claims empirically testable. Two specific instances are the extent to which lexical strength is a separate dimension of word knowledge and the extent to which explicit and implicit word knowledge are distinct. Below, I review two strands of measurement validation studies conducted in second language research.

Measurement Validation

Researchers relying on quantitative data to address research questions and test hypotheses need valid measurement tools. Validity, in its most general sense, is the extent to

which a measure measures what it purports to measure (e.g., Sireci, 2009). At the same time, validity should be conceptualized as the quality of a test that is relevant to the interpretation and use of its results (Messick, 1989). In this light, then, the validation process involves collecting evidence to corroborate or dispute these interpretations and uses (Messick, 1989). In a seminal paper, Messick (1989) foregrounded construct validity in his unified, single view of validity (as opposed to a multifaceted view of validity, such as content and criterion-related validity). In other words, when interpreting test results, to what extent do these interpretations relate to the construct in question? In order to focus on interpretations of test results, one needs to first understand the inference test users and researchers make from the assessment (Chapelle, 1999, 2021). Recently, Chapelle (2021) provided a list of seven types of inferences that are commonly made. The most relevant to the present context is domain definition. The question at hand is the extent to which the domain of vocabulary knowledge is "adequately analyzed to create test tasks that elicit relevant performance" (Chapelle, 2021, p. 16). As established above, certain aspects of the vocabulary construct are not well captured by traditional, offline tasks. In that sense, the limits on domain coverage of these tests cast doubts on the interpretations of the results (e.g., those scoring high have higher lexical skills) because, for example, lexical strength is under-represented (if at all). Now that time-sensitive words measures are available, researchers can expand the domain coverage of the vocabulary construct, and at the same time refine its definition and evaluate the extent to which these time-sensitive measures accurately summarize performance. However, to date, there has been a limited number of construct validation studies involving comparison between multiple measures in vocabulary research. Therefore, in this section, I will first review a somewhat

parallel literature in construct validation studies in grammar research before returning to validation work that has been conducted in the domain of vocabulary.

Measurement Validation Studies in Grammar Research

As mentioned in the section on implicit and explicit word knowledge, many grammar researchers differentiate explicit from implicit knowledge (e.g., Andringa & Rebuschat, 2015; DeKeyser, 2003; N. Ellis, 2005; Hulstijn, 2005). In other words, the construct of grammatical knowledge is believed to be two-dimensional in nature, at least according to these authors. To test this psychological dimensionality, researchers need measures that have a similar psychometric dimensionality (Henning, 1992). Put differently, researchers need separate, independent items and/or tests that tap into these two psychological dimensions (i.e., explicit and implicit knowledge). Caution is required when psychological and psychometric dimensionality do not correspond, potentially suggesting that certain dimensions cannot be measured or have not been theorized appropriately. Therefore, identifying this correspondence between psychological and psychometric dimensionality has been the theme of this line of validation studies (e.g., R. Ellis, 2005; R. Ellis & Loewen, 2007; Godfroid & Kim, in press; Gutiérrez, 2013; Spada et al., 2015; Suzuki, 2017; Vafaee et al., 2017).

In a seminal paper by R. Ellis (2005), the author administered a total of five tests, two of which were hypothesized to measure explicit knowledge, and three of which were designed to tap into implicit knowledge. In the operationalization of the explicit measures, participants made judgments on the grammaticality of sentences presented to them. In this grammaticality judgment task, no time pressure was imposed on the test takers. Participants also completed a meta-linguistic knowledge test where they identified and explained grammatical errors. For the

implicit measures, participants engaged in an oral production task and an elicited imitation task where they orally repeated spoken sentences (some containing grammatical errors) in correct English. Finally, in a separate grammaticality judgement task, they were placed under time pressure as they were responding to the stimuli. The results of these tasks were subjected to a principal component analysis which showed two underlying components of what had been measured by these tasks, aligning well with the initial hypothesis (i.e., the implicit and explicit measures tapping into implicit and explicit knowledge, respectively).

Researchers have followed up the findings reported by R. Ellis (2005) (e.g., Gutiérrez, 2013; Spada et al., 2015; Suzuki, 2017; Vafaee et al., 2017). They often administer a battery of tests and conduct a factor analysis on the test results to assess the correspondence between the psychological (e.g., implicit *vs*. explicit knowledge) and psychometric dimensionality of the battery (e.g., a two-factor solution in the factor analysis). For example, Gutiérrez (2013) adapted a subset of tasks from Ellis (2005), both timed and untimed written grammaticality judgement tasks and a meta-linguistic test. The author separately submitted the grammatical and ungrammatical items on the two judgement tasks (four measures) together with the meta-linguistic test to a confirmatory factor analysis. Results showed that the ungrammatical items from both judgement tasks and the meta-linguistic test loaded onto a factor the author labelled as explicit knowledge. The grammatical items, on the other hand, loaded on another factor called implicit knowledge. Based on these results, the author suggested that grammaticality of items can determine whether a task is an explicit or implicit measure. In addition, Gutiérrez (2013) found insufficient evidence that time pressure alone plays a similar role in a task in terms of measuring explicit versus implicit knowledge.

This line of work has been important to SLA researchers because it provides validity evidence for a particular (set of) task(s). Researchers interested in the development of a certain type of knowledge may take advantage of this literature when deciding on their own measures. In a study by Issa et al. (2020), for example, the authors administered an elicited imitation task and acceptability judgement tasks to measure grammatical development of their participants during study abroad. The validation ground laid by previous research allows Issa et al. (2020) to claim that they have covered key domains of grammar knowledge in their measurement. Another advantage of these validation studies is that they demonstrated how grammatical knowledge may be modelled at the latent (i.e., unobserved) level, essentially allowing for the measurement of theoretical constructs via real world measures and allowing for a test of whether or not real-world measures tap into the theoretical constructs. In addition, researchers can use multiple measures to distill a purer measure of grammatical knowledge by accounting for measurement errors in individual tests and the different degrees of strength between the measures and the latent construct (Brown, 2015). Methodological discussions on this literature have also highlighted the principled use of confirmatory analysis (R. Ellis & Loewen, 2007; Isemonger, 2007; Vafaee & Kachinske, 2019), informing subsequent researchers of the best statistical practices in this line of work. Building upon the review of validation studies in grammar, I return to two validation studies in vocabulary research.

Measurement Validation Studies in Vocabulary Research

Although validating vocabulary tests is not new (e.g., Schmitt et al., 2020), there have been only two published validation studies that focus on construct validity of a battery of tests in L2 vocabulary research: González-Fernández and Schmitt (2020) and Koizumi and In'nami

(2020). Using confirmatory factor analyses, these authors tested the extent to which different vocabulary measures can separately tap into different constructs under the word component knowledge approach to vocabulary depth (e.g., Schmitt, 2014). Their approach, therefore, mirrors that has been taken by grammar researchers seeking to identify a correspondence between psychological and psychometric dimensionality.

In González-Fernández and Schmitt's (2020) study, the authors were interested in how relationships between different word knowledge components should be conceptualized. In other words, to what extent are different aspects of word knowledge distinct? Participants were one hundred and forty-four Spanish learners of English. They completed a total of eight tasks, tapping into four word knowledge components (i.e., form-meaning link, derivatives, multiple meanings, and collocations) at two levels of sensitivity (i.e., recall and recognition). There were a total of 20 target words repeated across all the tests, which ranged from the first to the ninth 1000 most frequently occurring words (1K – 9K). The authors initially hypothesized a second-order model where the highest, second-order latent variable represents vocabulary knowledge in a general sense. The indicators of this latent variable were four first-order latent variables, representing the four word knowledge components. Each of these four first-order latent variables had two observed indicators (its respective recall and recognition tests). However, the model-implied variance-covariance matrix did not replicate the characteristics in the data, indicating that the hypothesized model was not an acceptable representation of the data. In other words, a correspondence between psychological and psychometric dimensionality was not found.

The author, then, revised the model specifications such that eight indicators (four word knowledge components at two levels of sensitivity) loaded to one factor named vocabulary knowledge. In addition, the authors added correlations between the residuals of the recognition and the recall tests for each word knowledge component. The authors reported a good fit for this model. Based on the factor loadings (.81 – .93), the authors suggested that "all word knowledge aspects mak[ing] a large contribution to the explanation of the Vocabulary Knowledge construct" (González-Fernández & Schmitt, 2020, p. 497).

Although the initially proposed multi-dimensionality of word knowledge based on different word knowledge components was not empirically supported, this study represented the first of its kind, and it has shed important light on this area of research. For example, depth was conceptualized only as word knowledge components (i.e., knowledge of more word components signaling greater depth of knowledge). As the authors alluded to, the relationships between recognition and recall can be further examined from a developmental perspective, essentially investigating the extent to which, for example, receptive and productive knowledge can be distinctly measured. This follow-up would then conceptualize the construct of depth as a developmental trajectory in line with what has been discussed previously (e.g., Schmitt, 2014). More generally, this approach of measurement validity can be applied to test some competing conceptualizations of vocabulary knowledge. For example, researchers can use this approach to address the long-standing relationships between vocabulary size and depth (e.g., Schmitt, 2014), which in fact was exactly the aim of Koizumi and In'nami's (2020) study.

In Koizumi and In'nami's (2020) study, 225 Japanese learners of English took a total of five vocabulary tests: one measuring vocabulary size and four tapping into vocabulary depth as

operationalized, again, as word knowledge components (i.e., word association, polysemy [L1 to L2 and L2 to L1], collocation). Words differed across different tasks but were all sampled from a wordlist compiled by a local teacher association (i.e., Japan Association of College English Teachers). The author further broke down the size test into three frequency levels to be submitted to the confirmatory factor analysis as different indicators. In total, then, there were seven indicators (three from the size test, four of each of the tests of depth). The one-factor model had a latent variable named size and depth, onto which all seven indicators loaded. This model represented a unified, single construct of vocabulary knowledge. The two-factor solution had correlated, but separate factors for size and depth. Three indicators from the size test loaded to the size factor, while the four tests of depth loaded to the depth factor.

In terms of the results, both the one- and two-factor solutions produced a good fit, indicating that they were both good representations of the data. The two-factor solution had a better fit as assessed by common fit indices (e.g., CFI, RMSEA, SRMR), AIC, and a chi-square difference test of deviance. These results suggested that the measures demonstrated a psychometric dimensionality that echoes the theorization of size and depth. Despite the evidence for distinct dimensions for size and depth, the correlation between the two at the latent level was .945. Since the use of confirmatory factor analysis takes measurement errors into account, this correlation estimate is more accurate than previous studies correlating individual size and depth measures (for a review, see Schmitt, 2014). On this account, then, this high correlation may point to a lack of practical significance to differentiate size from depth because of the lack of divergent validity.

Taken together, the two studies reviewed have investigated the dimensionality of vocabulary measures and have focused upon vocabulary depth operationalized as multiple word knowledge components. Appropriate for their studies, these authors have relied exclusively on explicit word measures. In this light, there is much room to investigate the extent to which other different conceptualizations as reviewed in the first section of this chapter (e.g., lexical strength and implicit word knowledge) can be measured in a psychometrically distinct manner. In other words, this line of work will shed important light on how well word measures can be mapped to the theoretical conceptualizations of vocabulary knowledge as a construct. For example, are traditional and time-sensitive measures of vocabulary psychometrically distinct? Should vocabulary strength (which underlies fluent use of language) be conceptualized as a separate, independent dimension (e.g., Daller et al., 2007; Harrington, 2018; Yanagisawa & Webb, 2020)? Are explicit and implicit word knowledge (e.g., Elgort, 2011; 2018) distinguishable at the behavioral level? As alluded to earlier, an additional advantage of this line of research is that it will also provide a solid basis for researchers to take advantage of multiple measures in modeling vocabulary knowledge at a latent level. In this regard, it will be useful to understand the predictive validity of the vocabulary construct under different conceptualizations.

The Present Study

In the light of the literature reviewed in this chapter, I conducted the present construct validation study of a battery of six vocabulary measures. There were two overall goals: first, to assess the alignment between the psychological dimensionality of word knowledge, as theoretically conceptualized differently by vocabulary researchers, and the psychometric

dimensionality of its measurement, as tested statistically in the present study; second, the examine the predictive validity of the vocabulary construct under different conceptualizations. This study will adduce empirical evidence to the construct validity of time-sensitive measures of vocabulary that are believed to be qualitatively different from traditional explicit tests. This evidence will complement the argumentation approach that researchers have taken to contrast (explicit) knowledge and lexical strength and implicit word knowledge (e.g., Godfroid, 2020b). The predictive validity evidence will also inform researchers what conceptualization may be more superior in terms of accounting for individual differences in general proficiency. I formulated the following research questions to guide the study:

RQ1a: To what extent do implicit word measures demonstrate a distinct psychometric dimension from explicit word knowledge measures?

RQ1b: To what extent do time-sensitive measures of lexical strength demonstrate a distinct psychometric dimension from untimed word knowledge measures?

RQ2a: How well can the vocabulary construct conceptualized as explicit and implicit knowledge predict self-reported general proficiency?

RQ2b: How well can the vocabulary construct conceptualized as word knowledge and strength t predict self-reported general proficiency?

CHAPTER 2: METHODOLOGY

In this chapter, I present the methodology to address the research questions for the present study. I detail the information on the participants, the critical words, the measures, and the procedure. At the end of the chapter, I also present the data analysis plan.

Participants

Given the current research aims, I engaged in sample size planning based on overall model fit of a hypothesized CFA model (e.g., K. H. Kim, 2005), as opposed to the power required to detect an effect (i.e., a significant regression path) (e.g., Muthén & Muthén, 2002). In an initially hypothesized, two-factor CFA model, I had six observed variables (degrees of freedom available = 6 (6+1)/2) = 21) and 13 freely estimated parameters. Although one measure was later dropped, which led to a revision of model specifications (see Measures and Data Analysis below), these numbers meant that the initially hypothesized model had eight degrees of freedom (21 - 13 = 8). I used this information for sample size planning prior to data collection. The formulas provided by K. H. Kim (2005) suggested that the desirable sample size ranged from 127 to 752, depending on the chosen fit indices, the strength of factor loadings (λ_x), and the desired fit and power levels.

Table 2 summarizes these recommended sample sizes based on some conventional criteria of model fit (e.g., < .05 for Root Mean Squared Error of Approximation [RMSEA] and > .95 for Comparative Fit Index [CFI], Hu & Bentler, 1999).

Table 2
Desired Sample Sizes Based on Model Fit

Power	RESEA	CFI	λ_x	Desired Sample Size (N)
.80	.05			752
.80		.95	.80	127
.80		.95	.60	429

Note. RMSEA = Root Mean Squared Error of Approximation; CFI = Comparative Fit Index; λ_x = Factor loadings in the factor analysis model

Given this range of sample sizes, one practical determining factor appeared to be the expected strength of factor loadings. I then consulted the two previous studies using a similar data analysis approach in this research field (i.e., L2 vocabulary studies): González-Fernández and Schmitt (2020) and Koizumi and In'nami (2020). In both studies, the factor loadings of the final model reported were above 0.80. At the same time, it was important to note that these studies only included accuracy-based measures. In addition to these measures, the present study incorporated a number of reaction time-based measures. These reaction time-based measures, according to L2 grammar research using a similar analytic approach, could have much lower factor loadings despite a satisfactory global fit (e.g., Suzuki, 2017). These low loadings could potentially result from relatively low reliability levels when experimental tasks are used to index individual differences between participants (e.g., Draheim et al., 2019; Rouder & Haaf, 2019). Therefore, it was less than straightforward to have an accurate *a priori* expectation of the factor loadings in the sample size planning stage. In this regard, this procedure also highlighted the difficulty for researchers to obtain sufficient, useful information for an accurate sample size plan, especially when a study is the first of its kind (e.g., Brysbaert,

2020). Considering (1) the sample sizes in González-Fernández and Schmitt (2020) (N = 144) and Koizumi and In'nami (2020) (N = 255), (2) the sample size range returned by the sample planning procedure based on model fit (see Table 2), and (3) practical considerations in terms of available funds and time (e.g., Loewen & Hui, 2021), I had planned for recruiting 150 participants.

In the end, one hundred and forty-five participants took part in the experiment. They were sampled from the international student population at Michigan State University. The participants were undergraduate and graduate students who majored in various disciplines and were speakers of a variety of first languages (L1), including but not limited to Mandarin Chinese (32%), Hindi (16%), Vietnamese (6%), Korean (4%), and Marathi (4%). Their demographic information, such as age and length of residence in the US, is presented in Table 3. As also noted by González-Fernández and Schmitt (2020), the analysis required a sample to have a reasonably large range of proficiency levels and hence sufficient between-participants variance in the data. Therefore, I strategically included, through different means of recruitment, participants of different levels of studies (e.g., freshmen, seniors, and MA and PhD students) and different lengths of residence in the US.

Table 3

Demographic Information About the Participants

	Mean (*SD*)	
Age	24.97 (5.37)	
Length of residence (in years)	3.16 (2.87)	
Frequency of English use (overall)[1]	6.44 (2.53)	
Frequency of English use (past week) [1]	6.71 (2.06)	
Frequency of English use (past month) [1]	6.79 (2.11)	
Self-rated proficiency[2]	7.0 (1.69)	
	Undergraduate	Graduate (e.g., MA, PhD or Professional Degrees)
Level of current study	44 %	56 %

Notes: [1] participants self-reported on a sliding scale where 1 represented "Never" and 10 meant "Always"; [2] participants self-reported on a sliding scale where 1 represented "Total beginner" and 10 meant "Native-like"

All participants received monetary compensation for their time. Ethical clearance was obtained from the Institutional Review Board (IRB) according to our university's regulations governing research involving human participants.

Critical Words

In choosing the critical words, I considered two factors: the number of words required and their frequency levels as appropriate for the participants (L2 learners in an American university setting). First, considering the number of critical words, I referred to both González-Fernández & Schmitt (2020) and Koizumi & In'nami (2020). I considered these studies to be relevant because they employed the same statistical analyses to address very similar research questions to the present research. In the study by González-Fernández & Schmitt (2020), the authors included 20 target words while Koizumi & In'nami (2020) had 20 to 40 items depending on the test. Since reaction time-based measures (see Measures below) typically have a low

level of reliability, Siegelman et al. (2017) suggested that researchers should increase the number of trials as a way to improve the ability of the tasks to discriminate individuals of varying ability levels. In addition, a relatively larger number of words would allow room for removal unsatisfactory items that demonstrate poor psychometric properties. At the same time, researchers need to avoid exploiting participants' time and effort. Having participants take part in an unnecessarily long experiment can be an ethical issue (e.g., Loewen & Hui, 2021). All considered, I decided to have initially 40 critical words.

In terms of the composition of the critical word set, I considered the expected vocabulary size of the sample and hence the expected variability in the data (i.e., the individual differences in participants' lexical proficiency). I took two pieces of information into account: first, non-native Ph.D. students are estimated to have a vocabulary of 9,000 word families (Nation, 2012). Second, with an inclusion criterion of a 9,000-word vocabulary size, Godfroid et al. (2018) reported their participants' vocabulary sizes to be between 9,100 and 12,200 word families as measured by the 14k Vocabulary Size Test (VST) (Nation & Beglar, 2007). Given that (1) my sample included both undergraduate and graduate students, (2) the VST is a receptive vocabulary test (i.e., easier than a productive task), (3) reaction time-based effects might be more difficult to detect, and (4) items in an experimental task used as an individual differences measure should vary across different difficulty levels (Siegelman et al., 2017), I decided to include words between the frequency bands of K2 and K5 as the critical words for the present study, representing a reasonably wide range with appropriate difficulty levels.

Table 4
List of Critical Words

2nd 1000 Level (K2)	3rd 1000 Level (K3)	4th 1000 Level (K4)	5th 1000 Level (K5)
maintain	soldier	compound	deficit
stone	restore	latter	weep
upset	jug	candid	nun
drawer	scrub	tummy	haunt
patience	dinosaur	quiz	compost
cap	strap	input	cube
pub	pave	crab	miniature
circle	dash	vocabulary	peel
microphone	poverty	remedy	fracture
pro	lonesome	allege	bacterium

After deciding on the frequency bands of the critical words, I had initially adopted the forty items between K2 to K5 from the 14k Vocabulary Size Test (Nation & Beglar, 2007), which was developed from the spoken section of the British National Corpus, to be the critical words for the present research. During piloting, I found that participants (N = 24) scored especially low for *nil* (33%) and *rove* (38%) on the Yes-No RT test (see details below), compared with the overall accuracy mean of 88% (*SD* = 32%) for the whole set of items. Therefore, I replaced these two words with *cap* and *poverty* both of which were in the same corresponding frequency bands. I present all 40 critical words in Table 4.

Data Collection Platform

All data were collected on Gorilla (www.gorilla.sc), an online platform which can be used for psycholinguistic research. The decision to collect data online was mainly due to the COVID-19 pandemic which resulted in (partial) closure of university buildings (hence our lab) and students leaving campus. Since online research, particularly in psychology, has only recently

grown, I considered a number of concerns researchers may have in relation to data quality (e.g., Woods et al., 2015). Here, I also outline measures that I implemented to mitigate some of these potential problems.

First, in terms of identity of the participants, one immediate reaction to online data collection can be that researchers have no ways to verify one's identity. In the present research, there were two key qualifying criteria for participants: (1) be a non-native speaker of English and (2) be a student at an American university. As in lab-based research, I relied on self-report by the participant of their non-native speaker status. For their student status, I requested that they provide a university email address for communication and for the Gorilla system to send a link to participate in the experiment. It is true that not all who possess a university email address is a student. But, I considered asking participants to provide further identification (e.g., student card) be unnecessary because that would mean collecting more personal information (e.g., student number and picture) than necessary for this study. Participants also self-reported their non-native speaker and student status twice: once in a screening survey, and once in Gorilla before the actual experiment (see Procedure below).

The second concern was potential attention lapses. In lab-based contexts, there may be a certain level of supervision on site from the researcher to maintain participants' attention. Such supervision is absent in virtual space. In the worst-case scenario, participants could be merely guessing, randomly responding to experimental items. In the data analysis procedure (see below), I paid close attention to individual participants' accuracy scores. Given the thoughts put into selecting the critical words, I expected participants to perform reasonably well. At the very least, they should perform above chance levels (50% accuracy when given two

forced choices and 25% when given four options in the multiple-choice format). In tasks that contained non-words (letter strings that do not form a word), I expected a low false alarm rate (incorrect responses to non-words, suggesting guessing) (e.g., < 50% in a binary, forced choice situation). I used these criteria to exclude participants who either did not pay attention or did not have the proficiency levels to provide useful information for this study. As for general fatigue, I arranged participants to take part in the experiment (of five tasks, see Measures below) on two separate days. Each day took approximately 30 – 45 minutes. I also built in breaks within and between tasks to allow participants to rest if needed.

Indeed, there has been a small literature comparing online and lab-based data collection (Crump et al., 2013; Germine et al., 2012; Klein et al., 2014; Ruiz et al., 2019). Together, these studies suggested that the two data collection settings do not differ to a considerable extent. For example, Germine et al. (2012) found similar results in their replication attempt using online tools to those reported in the initial studies. Importantly, the tasks involved in the study were considered to be more vulnerable to lapses in attention, such as the Cambridge Face Memory Test and the Forward Digit Span task. Similarly, comparable results obtained from the two data collection settings were reported by Crump et al. (2013) and Klein et al. (2014) who incorporated both reaction-time and memory tasks in their study. In second language acquisition research, Ruiz et al. (2019) also reported similar findings between the lab- and web-based versions of their working and declarative memory tasks.

While some of these findings may seem encouraging, the only task that Crump et al. (2013) failed to replicate (one in eight tasks) was a masked priming task involving symbol (i.e., arrows pointing to different directions). The general idea is that participants were expected to

respond faster to an arrow preceded by another one pointing to the same direction than when it was preceded by one pointing to a different direction. The authors suggested that experiments involving brief presentations of elements (e.g., 16, 32 ms) can be less reliable in internet-based research. In fact, Hamarick (2020) outlined a number of challenges face researchers using reaction-time based measures, even in lab-based settings. These challenges can influence the data quality to varying degrees depending on factors such as equipment, experimental set-up, and instructions to participants, all of which could introduce random variability (noise) in the data which in turn buries important signals that researchers look for (e.g., an expected effect of 20-ms difference in reaction time). The key questions at hand were the accuracy and precision levels of timing measurements of online data collection platforms and the extent to which such levels would be acceptable.

One study that systematically evaluating the accuracy and precision of online experiment platforms was Anwyl-Irvine et al. (2020). The authors investigated the impact of different system set-up combinations (platforms [e.g., Gorilla], browsers [e.g., Google Chrome], and operating systems [e.g., Windows on a laptop]) on display time across 30 different time frame durations and on reaction time recording. Results showed that, for example, Gorilla had a mean of 13.44 ms of visual duration delay with a standard deviation of 15.41 ms. It means that when a stimulus was due to be presented, it is only after, on average, 13.44 ms later that it was actually shown on the screen. It ranked third in terms of absolute mean values among the four platforms compared (mean values for delay ranged from -6.24 ms [stimulus presented before it is due] to 26.02 ms). Note that, if the delay had been consistent, it would have potentially posed less of a problem. However, the standard deviation of 15.41 ms indicated a

rather large variability given the scale. The more important measure was reaction-time recording. Gorilla recorded reaction times as the time between the actual presentation of the stimulus and a response. In a way, then, any (variable) delay of the presentation would not impact the reaction time recording. For reaction-time recording, the mean delay for Gorilla was 78.53 ms with a standard deviation of 8.25 ms. It means that the platform only detected a response by the robot actuator only after, on average, 78.53 ms the key was struck. Again, it ranked the third in absolute mean values, compared with other systems. Note that, however, the standard deviation was rather small, especially when compared with other platform whose standard deviations ranged from 15.27 ms to 28.16 ms. In other words, potentially, although there was a general delay, such delays can be relatively consistent on Gorilla, at least compared with other systems tested in the study.

Although the authors optimistically concluded that these platforms provided "reasonable accuracy and precision" (p. 1) in terms of display duration and reaction-time recording, there was still variability associated with what equipment the participant used. Also, in the context of the present study, some of the delays and variability in the delays represented potential random noise in the data which appeared larger than ideal. For example, for the priming tasks (see Measures below) which were most susceptible to reaction time accuracy and precision, I expected a group-level difference of 22 ms to 80 ms as informed by Elgort (2011). The potential random variability described above (e.g., a standard deviation of 8.25 ms in reaction time recording delays) may then prevent the signal (i.e., expected effects of priming) to be observed. In a way, then, the situation called for strategies to reduce the amount of noise in the data to the extent that was possible, which in turn meant that the signal could be emerge

more clearly (e.g., Siegelman et al., 2017). To achieve this, I engaged in item analyses to identify items that elicited random performance given the current experiment set up. As will be detailed in the data analysis section, removing items that did not elicit priming may render the task a more reliable measure (Siegelman et al., 2017). Furthermore, when random variance was inevitable, I attempted to incorporate such variability in the statistical modeling so that it was properly accounted for (Rouder & Haaf, 2019). Using mixed-effects modeling, variability due to participants (which resulted from both difference in their ability and technological set up) and items can be partitioned to highlight the expected effect (Rouder & Haaf, 2019) (see more discussion in the Data Analysis section).

Measures

In this study, there were a total of five tasks, all of which tap into the participants' form-meaning link of the critical words. I modeled these tasks after previous research in order to better align the present study with the existing research base and practices. These tasks were a form-meaning receptive task, a form-meaning productive task, a simple lexical decision task (or a Yes-No RT test), a masked repetition priming task, and a semantic priming task. For each task, I describe the knowledge intended to be measured, the task itself, and methodological details related to the administering of the test. I will also detail information obtained from the task construction and piloting stages to demonstrate the quality of the instruments. In Table 5, I summarize these tasks, their expected effects, constructs being measured, and the type of data they afforded for the final data set.

Table 5

Summary of Measures for the Present Research

Task	Expected Effect	Test Construct	Explicit or Implicit	Timed or Untimed	Outcome Variable
Form-Meaning Receptive Test	NA	Knowledge of the form-meaning link at the level of recognition	Explicit	Untimed	Accuracy data
Form-Meaning Production Test	NA	Knowledge of the form-meaning link at the level of recall	Explicit	Untimed	Accuracy data
Yes-No RT Test	NA	Access to the form-meaning link	Explicit	Timed	Accuracy and reaction time data
Masked Repetition Priming Task	Repetition priming (faster responses to identity prime-target pairs)	Lexical representation	Implicit	Timed	Reaction time data
Semantic Priming Task	Semantic priming (faster responses to prime-target pairs that are semantically related)	Semantic representation	Implicit	Timed	Reaction time data

Note. NA = Not Applicable

Form-Meaning Receptive Test

I adopted the thirty-eight items in the K2 to K5 frequency bands from the 14k Vocabulary Size Test (Nation & Beglar, 2007) for this task. For *cap* and *poverty,* the two critical words replacing the original, unsatisfactory items identified during piloting, I wrote the test items myself as an experienced teacher of English as a foreign language. This task was designed to measure participants' knowledge of the form-meaning link at the sensitivity level of meaning

recognition. Items were clustered according to their frequency levels. At each level, there were ten items. For each item, I presented the target word together with a sentence in which the target word was used, as well as four options of definitions. Only one of these options was correct in describing the sense of the target as used in the sentence. The task for the participant was to choose the closest meaning to the target word (see an example item below and the full set of the test in Appendix A. The definitions for the target were always of higher frequency levels (i.e., more common) than the critical word in question. This manipulation was to minimize the probability that the participant would fail to select the answer because they did not know the words in the definitions. Here, I present an example item for the critical word *patience*:

PATIENCE: He has no **patience**.

a. will not wait happily

b. has no free time

c. has no faith

d. does not know what is fair

I invited three native speakers of English to attempt the test although the 14k Vocabulary Size Test (Nation & Beglar, 2007) had been used somewhat widely in vocabulary research (Godfroid et al., 2018; Peters, 2019; Vafaee & Suzuki, 2020) and had been subjected to

psychometric validation (Beglar, 2010). The native speakers all obtained a perfect score. In terms of assessing the face validity of the test, participants needed to have established the form-meaning link of the critical word in order to select the correct meaning. Also, the sentence was written in such a way that guessing meaning from context was not possible. Finally, the definitions were presented as choices, hence participants only needed knowledge at the sensitivity level of recognition to complete the task. Taken together, I consider this test to have sufficient face validity as an explicit, untimed measure of the form-meaning link at the level of recognition (i.e., receptive knowledge of the form-meaning link).

Form-Meaning Productive Test

I modeled the format of this test after the productive Vocabulary Levels Test (Laufer & Nation, 1999). This test was designed to measure the participants' productive knowledge of the form-meaning link. In particular, it tested participants' "controlled productive ability" (Laufer & Nation, 1999, p. 36) in the sense that learners needed to produce the target words when given a meaningful, obligatory sentence context. As an experienced teacher of English as a foreign language, I wrote one item for each of the forty critical words. For each item, I supplied a meaningful sentence context as well as the first letter(s) of the critical word. The provision of the first letter(s) was to prevent learners from filling in other legitimate alternatives. In writing these items, I referred to sentence examples in dictionaries, such as the Collins Dictionary (www.collinsdictionary.com). Care was given not to include words that are of lower frequency in the sentence context than the critical word. This was to minimize cases where participants' failure to supply the target was because they could not understand the context. Here is an example item for the target word *patience* (see the full set of items in Appendix B).

In the end, I lost my pat_____ and shouted at them.

To gather validity information, I engaged with four rounds of revision with four native speakers of English who were writing consultants at the University's Writing Center. In each round, the native speaker attempted the task. For items that were not attempted correctly and/or were deemed confusing, we discussed ways to modify the context and/or wording to improve the items. After each review, I revised the items according to their feedback and proceeded to the next round of review with another native speaker. In the final, fourth round, the native speaker was able to provide correct answers to all items. In the design of this test, participants needed to have productive knowledge of the target's form-meaning link (i.e., spelling and meaning). Although this assumed understanding of the sentence context, I considered this test to be valid for tapping into such explicit knowledge in an untimed manner.

Yes-No RT Test (Access to the Form-Meaning Link)

I modeled this task after previous studies that implemented a computerized version Meara's (2010) Yes-No test (e.g., Hui & Godfroid, 2020; Pellicer-Sánchez & Schmitt, 2012). This task was designed to measure access to the form-meaning link under time pressure (i.e., lexical strength). Stimuli included the 40 critical words for the present study. In order to make it a genuine task for participants, another 40 non-words obtained from the ARC Nonword Database (Rastle et al., 2002) were also included. Although there has been little consensus on the proportion of non-words, the typical range falls between 25% and 50% (Beeckmans et al., 2001;

Pellicer-Sánchez & Schmitt, 2012; X. Zhang et al., 2020). I chose 50% (i.e., there was one non-word for every real word) in an attempt to reduce guessing. In total, there were then 80 trials (40 real words and 40 non-words). Participants were asked to indicate if they knew a given word. To be more specific about what it means to know a word for this task, I followed Pellicer-Sánchez and Schmitt (2012), who told participants that a yes response means that the participant would recognize the word in a text and know its meaning(s). The participants indicated their knowledge by pressing the corresponding buttons on their keyboard (the j key representing Yes and the F key representing No). I instructed participants to judge as quickly and as accurately as possible, following conventions in psycholinguistics to guide participants to place equal emphasis on response accuracy and speed so that the accuracy and reaction time data can manifest performance differences somewhat equally (e.g., Draheim et al., 2019).

Each trial started with a fixation cross (+) presented for 400 ms, which was followed by the target in lowercase (e.g., patience) until the participant responded. The next trial followed after 100 ms. All trials were randomized by Gorilla. There was a practice block of six items at the beginning. Feedback was provided only in the practice block. All stimuli are presented in Appendix C. This task afforded both accuracy and reaction time data.

I consider this task to be a valid test of acess to the form-meaning link because participants needed to know the word meaning in order to respond correctly and that time pressure was imposed on them. In terms of the reaction time data, I made the assumption that "more hesitant and inaccurate responses would be slower, whereas more certain and accurate ones would be faster" (Pellicer-Sánchez & Schmitt, 2012, p. 492-3). Therefore, shorter reaction times was taken as evidence of more efficient access to the form-meaning link, a manifestation

of lexical strength. In other words, I used this task as a measure of lexical strength in the present study. In addition, since the task induces participants' awareness of the knoweldge assessed, this test is an explicit test.

I piloted this task with 24 learners of English drawn from the target population. As mentioned in the Critical Words section, the overall accuracy for real words was 88% with a standard deviation of 32%. The by-participant analysis suggested that participants performed generally well with real words. Accuracy ranged from 0.70 – 1.00 with a median of 0.93, suggesting ceiling effects for some participants. However, there was some guessing as demonstrated in the non-word data. The average false alarm rate (incorrect yes responses to non-words) was 12% with a standard deviation of 33%. The false alarm rate in the by-participant analysis ranged from 0.00 to 0.55 with a median of 0.05. These results underscored the need to account for guessing in the data analysis (Huibregtse et al., 2002; X. Zhang et al., 2020) because some participants guessed more than 50% of the time.

The by-item analysis revealed that two words had a very low accuracy (0.33 and 0.38), which I reported above in the Critical Words section. These two words were then replaced. The remaining word items showed a satisfactory accuracy range from 0.71 to 1.00. There were ceiling effects for 12 items to which all participants responded correctly (accuracy = 1.00), meaning that these items had no discriminant ability for this sample. For the sake of computing reliability for this pilot test, I removed these items because they demonstrated no item variance. With the resulting data set, Cronbach's alpha was .77, and McDonald's omega was also .77. In order to maintain test equivanlency across the five tasks for this study, I kept the items with ceiling effects in the stimulus list.

In terms of reaction time data, I only examined the correct yes responses. I trimmed items for which the reaction time fell outside the 300 ms - 2500 ms window (e.g., Jiang, 2013). Table 6 summarizes the reaction times in this pilot test. Overall, participants responded descriptively faster to real words than to non-words, which was expected (e.g., Scarborough et al., 1977; Stenneken et al., 2007). Analyzing only real word data, the split-half relability was .85. Taken together, the pilot results suggested this test worked as intended.

Table 6
Means and Standard Deviations of Reaction Times for Words and Non-words

Mean RT in Millisecond (*SD*)	
Real words	Non-words
778 (295)	846 (290)

Masked Repetition Priming (Lexical Representations)

I modeled this task after the operationalization of the masked repetition priming paradigm in Elgort (2011). As discussed in the Literature Review, this task taps into participants' lexical representations. When the lexical representations in question are established in the mental lexicon, participants are expected to make a faster lexical decision when the target is preceded by an identical prime than by an orthographically or semantically unrelated prime. This facilitation (priming) can only be observed when the target is lexically represented in memory. The general idea is that the word prime pre-activates the lexical representation in question, and so access to it is easier (faster) when the participant sees the target, resulting in a faster response. In contrast, in cases where no lexical representations are established in

memory, the prime will have no effects on the response to the target because there is no pre-activation in the absence of lexical representation.

In this task, each critical word constituted an item. Each item was a duplet (i.e., consisted of two trials). One trial was in the related condition and the other in the unrelated condition. In the related condition, the prime and the target were identical (e.g., patience–PATIENCE). In the unrelated condition, the prime was changed such that it was not related to the target in form (i.e., no letters are in the same position of the words) and meaning (e.g., occasion–PATIENCE). I matched the prime in both conditions by length, parts of speech, frequency in Zipf (according to Brysbaert & New [2009] with a tolerance of +/- 0.15), and character bigram probability (according to Brysbaert & New [2009] with a tolerance of +/- 0.001) using the LexOPS package in R (Taylor et al., 2020). The task for the participant was to make a lexical decision on the target (e.g., to judge whether or not PATIENCE was a word in English). With 40 critical words, there were 80 critical trials (one trial in each of two conditions). Following Elgort (2011), I further reduced the proportion of related trials in the stimulus list by introducing an additional 80 unrelated word pairs as fillers. This was to minimize the prime validity effect, which was priming due to a high proportion of repetitions (e.g., Bodner & Masson, 2001). To make the lexical decision a genuine task for the participant, there were 160 non-word trials, half in the related (repetition) condition and half in the unrelated condition. Table 7 provides a summary of different trial types. All stimuli are presented in Appendix D.

Table 7

Summary of Trial Types in the Masked Repetition Priming Task

Trial Type	Condition	Prime	Target	Target Type	No. of Trial
Critical	related	patience	PATIENCE	critical words in the study	40
Critical	unrelated	occasion	PATIENCE	critical words in the study	40
Filler	unrelated	brother	SONG	other real words	80
Nonword	related	snarbs	SNARBS	non-words	80
Nonword	unrelated	plisc	SNARBS	non-words	80

In total, there were 320 trials, of which 200 (63%) were unrelated and 120 (37%) were related. Participants took part in all trials. Following Elgort (2011), I used the standard three-field masking paradigm. This means that, following a fixation cross (+) presented for 400 ms, I first presented a forward mask (a string of hash signs [###]). This mask stayed on the center of the screen for 500 ms. Immediately after that, the prime in lowercase (e.g., patience) appeared for 55 ms in the same space as the mask, followed by the target in uppercase letters (e.g., PATIENCE) for 500 ms before turning into a blank screen. This blank screen was displayed until the participant responded. The next trial started after 100 ms. The use of both lower- and upper-case was to ensure that the participant would respond to the target (not the prime) and to rule out the effect of visual overlap between the two. In the present case of a repetition priming task, different capitalization made it clearer to the participant what the target was although participants were not expected to be aware of the presentation of the prime. I asked participants to judge whether or not the presented letter string (the target) forms a word in English (i.e., to make a lexical decision). As in the Yes-No RT test, participants made their judgements by pressing the corresponding key on their keyboard. I also instructed participants to judge as quickly and as accurately as possible. There was no feedback, except in the practice

block of six items. Breaks were inserted approximately every 80 trials to avoid fatigue. All trials were pseudo-randomized by Gorilla. This task afforded reaction time data from which I analyzed the level of priming (facilitation) as a measure of implicit lexical-formal knowledge of the critical words.

I piloted this task with a sample of 24 participants drawn from the target population. The overall accuracy rates for all trials (words and non-words), word trials (critical words and fillers), and critical word trials were 81% (*SD* = 39%), 85% (*SD* = 36%), and 88% (*SD* = 36%), respectively. The false alarm rate (incorrect responses to nonwords) was 22% (*SD* = 42%). From the accuracy data, then, it appeared that participants generally had the proficiency to complete the task, consistent with the pilot results of the Yes-No RT test. However, there was a certain level of guessing with a rather large variability between participants, pointing to the needs to consider individual participants' false alarm rate to ensure data quality.

To demonstrate that this task was able to generate the intended the repetition priming effect as an indication of established lexical representations, I focused on two aspects of the reaction time data: first, for the critical trials, participants were expected to respond faster to the target in the related trials than in the unrelated trials (i.e., there should be facilitation as a result of repetition), and second, such priming would not be observed in the non-word data. As with the Yes-No RT test, I trimmed the reaction time data using the lower and upper thresholds of 300 ms and 2500 ms, respectively (Jiang, 2013). I then computed descriptive statistics of the reaction times in both conditions by trial types (see Table 8).

Table 8

Means and Standard Deviations of Reaction Times for Critical Words Between Conditions (Repetition Priming)

	Mean RT in Millisecond (*SD*)	
	Related	Unrelated
Critical words	(e.g., patience – PATIENCE)	(e.g., occasion – PATIENCE)
	623 (222)	650 (209)
Non-words	(e.g., rects – RECTS)	(e.g., flief – RECTS)
	720 (197)	718 (196)

Elgort (2011) reported a 52-ms priming for the target words in her study (i.e., pseudowords that her participants had learned recently) and a 75-ms priming for the real, low frequency word trials. In the present pilot data, I obtained a 27-ms priming effect, which was consistent with but somewhat smaller than the previous results. I followed up on this 27-ms difference with mixed-effects modeling, an appropriate statistical approach to handle nested data in psycholinguistic experiments (Baayen et al., 2008). Nested data call for the incorporation of cross random effects (i.e., each participant provided multiple, correlated data points and each stimulus elicited multiple observations that autocorrelate) (e.g., Baayen et al., 2008). This simultaneous handling of random effects due to participants and stimuli represents a key advantage of the technique over separate by-participant and by-item analyses where data are aggregated, leading to loss of statistical information. In building the mixed-effects models, the outcome was always a reciprocal transformation of reaction time (i.e., -1/RT). I started with a null model with no fixed effects but only the random intercepts by participants and by items. This null model provided an intra-class correlation of .38 as statistical evidence for the need to account for the random effects. Then, I added condition (related [0] *vs.* unrelated [1]) as a fixed effect. In terms of the random effect structure, I built a maximal model where all random

intercepts and slopes were entered (Barr et al., 2013). The significance of condition indicated that the 27-ms difference was reliable, after controlling for random variability between participants and items. Hence, the current task successfully elicited the targeted priming effects from participants. The model summary is presented in Table 9.

Table 9

Summary of Mixed Models for Pilot Data - Masked Repetition Priming Task

	m0 (null model)		m1 (maximal)	
Fixed effects	estimate [a] (*SE*)	*t (p)*	estimate (*SE*)	*t (p)*
Intercept	-1.67 (0.05)	-33.23 (<.001)	-1.73 (0.06)	-30.45 (<.0001)
Condition			0.11 (0.02)	4.80 (<.001)
Random effects	Variance (*SD*)	Intercept-slope correlation	Variance (*SD*)	Intercept-slope correlation
By-participant intercept	0.05 (0.24)		0.07 (0.26)	
By-item intercept	0.007 (0.08)		0.01 (0.10)	
By- participant slope for condition			0.004 (0.07)	-.77
By- item slope for condition			0.003 (0.06)	-.83
Residual	0.10 (0.32)		0.10 (0.31)	
AIC	-22161		-22219	

Note. [a] all estimates were multiplied by 1000 for easier reading.

I also conducted by-participant and by-item analyses where reaction times were aggregated across items and participants, respectively. In the by-participant analysis, 18

participants (out of 24) showed a net positive difference between the related and the unrelated conditions (unrelated minus related), indicating some priming at the descriptive level. These differences ranged from 10 ms to 171 ms. The remaining six participants showed a net negative difference, ranging from -6 ms to -72 ms. In the by-item analysis, 29 items (out of 40) showed a net positive difference between the related and the unrelated conditions (unrelated minus related), indicating some priming at the descriptive level. These differences ranged from 0.2 ms to 190 ms. The remaining 11 items showed a net negative difference, ranging from -2 ms to -161 ms. These analyses showed that despite an overall effect of priming, there remained a certain amount of variability between participants and between items, highlighting the need for participant- and item-level inspections.

I then proceeded to analyzing the non-word data. As mentioned, no priming was expected in the non-word data, because participants should not have these letter strings represented in memory as lexical items. Therefore, the prime should not facilitate the response to the target even when it was repeated. From the descriptive statistics (see Table 8), there was a 2-ms difference between the conditions. In the light of the standard deviations, I concluded that there were no reliable differences between the two conditions, supporting the lexical nature of masked repetition priming.

Finally, in order to assess the extent to which participants were aware of the prime, I asked participants in a debriefing survey progressively (1) whether or not they had seen something between the fixation cross and the target, and (2) if so, what they had seen. All participants mentioned the mask (i.e., the hash signs), and no participants reported seeing a word (i.e., the prime). Therefore, the masking of the prime in this task worked as intended,

lending further support to this task as an implicit measure of vocabulary knowledge. Taken together the accuracy and the reaction time analyses, this masked priming task appeared to function as intended at the piloting stage.

Semantic Priming (Semantic Representations)

I modeled this task after Elgort's (2011) operationalization of the semantic priming paradigm to measure the establishment of semantic representations of words in memory. The general idea behind the priming effects expected in this task is similar to that in masked repetition priming task described above. Specifically, participants were expected to respond faster to a related prime-target trial (e.g., patience–calm) than an unrelated one (e.g., chestnut–calm). This facilitated processing of the target (e.g., calm) can be attributed to the prime pre-activating the overlapping semantic representations of the target. On the other hand, when the prime and the target are unrelated, the activation of the representations associated with the prime does not spread to those associated with the target because prime and target are not meaningfully related or interconnected. Therefore, I made the assumption that if learners show priming in this implicit task, they have integrated the prime into their semantic network of both the prime and the target in memory.

In this task, there were 40 critical items. For each item, there were two trials (i.e., a duplet), and each trial appeared in one of two conditions: semantically related and unrelated. In the related condition, the prime and the target were related semantically (e.g., patience-calm), and they were not related in the unrelated condition (e.g., chestnut-calm). The critical words of this present study were primes in the related trials while a matched prime was chosen

for the unrelated trials. That way, participants responded to the same target (e.g., calm) across conditions, allowing a meaningful comparison.

In establishing semantic (un)relatedness, I relied on the web application Snaut, a platform for semantic association evaluation (Mandera et al., 2017). Specifically, I obtained the cosine distance value for each critical trial, which represents the semantic space between the prime and the target (Günther et al., 2016). The lower the value, the smaller the semantic space there is between the two (i.e., the more related). At the item level, the related trial always has a smaller value than the unrelated trials with a mean difference of 0.31 units. At the group level, an independent Welch *t*-test was performed to test the significance of the difference. The assumption of equal variances was violated (Levene's test: $p = .02$), but the assumption of normality in both groups held (Shapiro-Wilk test: $p = .071$ and .072). Results confirmed that the cosine values for the related trials ($M = 0.62$, $SD = 0.11$, 95%CI [0.59, 0.65]) were statistically smaller than those in the unrelated trials ($M = 0.93$, $SD = 0.07$, 95%CI [0.90, 0.95]), t (37) = -15.14, $p < .001$, $d = -3.39$. This indicated that the semantic space between the related pairs was smaller (semantically closer) than that between the unrelated pairs, providing objective evidence for the validity of the manipulation.

I also took care to ensure that any priming to be observed would be due to the semantic relatedness, and not merely word association between the prime and the target. To measure word association, researchers often rely on databases created from word association tasks, where participants are presented with a stimulus and asked to give the first word that comes to their mind. It is important to note that there can be more than one reason for a particular response to come to the participant's mind. For example, words can be associated because

they often occur together (e.g., new – year). They can also be semantically related (e.g., synonyms). Therefore, in order to strengthen the internal validity of this semantic priming tasks, I followed Elgort (2011) in keeping the associative relationship between the prime and the target low. As a result, the task tapped into participants' semantic representations to the largest extent possible, rather than their associative knowledge of the critical words.

I used a web-based platform (http://rali.iro.umontreal.ca/word-associations/query/) to examine word association in the critical, semantically related trials, according to the Edinburgh Associative Thesaurus (EAT) (Kiss et al., 1973) and University of South Florida Free Association Norms (USF-FAN) (Nelson et al., 1998). I only used the forward association information because it reflected the order of presentation in the present task. For example, given *patience* (a critical word in the present study), the databases returned 18 tokens of 16 unique responses with the frequency of occurrence ranging from 1 – 2 (e.g., *tolerant*, *waiting*). This means that, for example, when given *patience,* two participants responded *tolerant.* From this information, I was able to estimate and quantify the strength of forward association between *patience* and *tolerant* (2/18 = 0.11). In the case of the *patience-calm* pair for the present study, no association between these two words is listed in both databases. Overall, 20 of the 40 critical, related pairs (50%) fell into this category. In 11 other pairs, the association was low (< 0.10). For seven pairs, the association was moderate (0.10 – 0.40). There were three pair (*latter-former; haunt-ghost; fracture-break*) that showed high association (>.40).

In addition, I made sure that the frequency of the target (e.g., *calm*) was of the same or a higher frequency band than the prime word (the critical word in the study, e.g., *patience*) so that knowledge of the target can be assumed. However, in three prime-target pairs, the target

belongs to a lower frequency band than the prime word. They were *strap-bra, dinosaur-fossil,* and *vocabulary-grammar*. Given the multiple dimensions of control (semantic relatedness, association, and frequency), these exceptions to the general rules for stimulus selection were considered as necessary concessions for the present study. At the same time, as detailed below, I attempted to control for variability specific to individual items through statistical means. If these exceptions had had an effect on the results, they would have been accounted for by the model. I present this association information in Appendix E and the full set of materials in Appendix F.

On top of the 80 critical trials, I added 80 unrelated filler trials to decrease the proportion of related trials in the stimulus list to minimize task taking strategies. I also introduced 80 non-word pairs which were presented twice. These 160 trials balanced the word and non-word trials to make it a genuine task for the participant. In total, then, there were 320 trials. Following Elgort (2011), I presented these trials in a list-wise fashion where participants made a lexical decision for each stimulus. Following a practice block of six items, the presentation list started with a fixation cross presented for 2000 ms. Then, each trial began with a blank screen (200 ms) followed by the stimulus (prime or target). In Table 10, I present a summary of the trial types.

Table 10

Summary of Trial Types in the Masked Semantic Priming Task

Trial Type	Condition	Prime	Target	Target Type	No. of Trial
Critical	related	patience	calm	target words in the study	40
Critical	unrelated	chestnut	calm	target words in the study	40
Filler	unrelated	brother	song	other real words	80
Nonword	unrelated	plisc	snarbs	nonwords	160

This task went through five rounds of piloting with native and non-native speakers of English. This process resulted in identification of programming errors, revisions of the prime-target pairs, and a change in the presentation method, all of which are reflected in the description above. Here, I report the last round of piloting involving 18 native speakers and ten participants drawn from the target learner population. These numbers were lower than the number of pilot participants in the first round as reported above. Although this was not optimal, it has also been suggested that a pilot study with a small number of participants can still be informative (Jiang, 2013). Specifically, the author suggested that "[a] general rule of thumb is that you can check the results after you have tested six to seven participants on each presentation list" (Jiang, 2013, p. 31). While Jiang (2013) acknowledged the low statistical power as a result of such a small sample, the author argued that researchers can gain insights into the direction and magnitude of an effect. For example, he wrote that "[running more participants] won't turn a weak -2 ms negative priming effect into a strong +34 ms positive priming effect" (Jiang, 2013, p. 31). Given the resources available, I took advantage of the information provided by these participants.

In the learner data, the overall accuracy rates for all trials (words and non-words), word trials (critical and filler trials), critical trials (prime and target trials), and target trials were 84%

(*SD* = 37%), 84% (*SD* = 37%), 89% (*SD* = 32%), and 92% (28%). The false alarm rate (incorrect responses to non-words) was 16% (*SD* = 36%), with one participant at 51%. In addition, it might be note-worthy that the accuracy for the prime trials (*M* = 86%, SD = 35%) was descriptively lower than the target trials (*M* = 92%, *SD* = 28%). This result made sense because the prime words in this experiment were the critical words for the study (e.g., *patience*), and the target trials were a semantically related word of a higher or a similar frequency level (e.g., *calm*). The high accuracy of the target trials also suggested that it was reasonable to assume participants' knowledge of the targets. Finally, the accuracy rates for the target trials by condition (i.e., preceded by a related *vs.* unrelated prime) were similar (91% [*SD* = 0.28] *vs.* 92% [*SD* = 0.27]). Taken together, these figured signaled that participants generally had the proficiency levels sufficient to complete task and that there were some levels of guessing, consistent with the two other reaction time-based measures.

In a list-wise presentation where participants respond to all stimuli (both the prime and the target), there is one additional factor to consider in data preparation: For the current semantic priming measure to be valid, participants need to know (respond correctly to) both the prime (in order to pre-activate the semantic representation) and the target. This general principle applies to both conditions (related *vs.* unrelated). In other words, in the data preparation, all four responses (to the prime and the target in both conditions) need to be correct for a given item of a given participant to be included for analysis. Although this could be a high bar, this requirement ensured the data quality in the reaction time analysis. As with the other reaction time tasks, reaction times were trimmed using a lower threshold of 300 ms and an upper threshold of 2500 ms. As a result of these criteria, one participant, who had a 51%

false alarm rate, had no data left in the resulting data set, other participants had 16 – 37 items (out of 40 items) in the data set. Different items had data for 1 to 9 participants (out of 9). In Table 11, I present the descriptive statistics for reaction times in both conditions.

Table 11

Means and Standard Deviations of Reaction Time for Learners in the Semantic Priming Task - Piloting

Mean RT in Millisecond (*SD*)	
Related (e.g., patience – clam)	Unrelated (e.g., chesnut – calm)
601 (170)	600 (164)

Although there was a lack of priming effects in either direction at the group level, I also inspected the by-participant analysis. Out of the nine participants, six showed a net positive difference in reaction times in the expected direction, ranging from 3.24 ms to 33.24 ms. For the other three participants, they showed a net negative difference (reverse priming), ranging from -4.51 ms to - 86.97 ms. In Elgort (2011), the author reported a 22-ms difference for her target words (i.e., pseudowords that her participants had recently learned) and a 37-ms difference for the real, low frequency word trials. Using Elgort's (2011) results as a reference, four participants (out of nine) showed signs of priming.

I also conducted a by-item analysis although each item had at most nine participants' reaction times to aggregate from, potentially making the results less reliable. Out of 40 items, 23 showed a net positive difference in the expected direction, ranging from 4.17 ms to 122.64 ms. The remaining 17 items had a negative reaction time difference ranging from -1.66 to -163.30. Using the figures reported by Elgort (2011), 15 (out of 40 items) showed signs of

eliciting priming. I did not engage in inferential statistics because of the sampling errors associated with only nine participants.

These results for learners were not optimal, but they were somewhat expected upon reflection. First, based on Elgort (2011), the expected effect size of the semantic priming task (i.e., a 22-ms difference) was smaller than that in the masked repetition priming task (i.e., a 50-ms difference), making it more difficult to detect any semantic priming reliably. Online data collection also inevitably introduced random variability to the data, worsening the situation. Finally, the sampling errors associated with the small sample size for this piloting could have prevented any trustworthy signals from emerging.

In the native speaker data, the overall accuracy rates for all trials (words and non-words), word trials (critical and filler trials), critical trials (prime and target trials), and target trials were 86% (*SD* = 34%), 90% (*SD* = 30%), 93% (*SD* = 25%), and 97% (18%). The false alarm rate (incorrect responses to non-words) was 21% (*SD* = 41%), with one participant at 51%. Again, the accuracy for the prime trials (*M* = 90%, SD = 30%) was descriptively lower than the target trials (*M* = 97%, SD = 18%), confirming again that it was reasonable to assume participants' knowledge of the targets. In terms of the reaction times, I present the descriptive statistics in both conditions in Table 12.

Table 12

Means and Standard Deviations of Reaction Time for Native Speakers in the Semantic Priming Task - Piloting

Mean RT in Millisecond (*SD*)	
Related (e.g., patience – clam)	Unrelated (e.g., chesnut – calm)
554 (161)	578 (168)

The mixed-effects model with the maximal random effects structure suggested a significant fixed effect of condition, indicating that the native speakers showed reliable priming at the group level. I present the model summary in Table 13. I took this result as evidence that the experiment was properly set up and that the manipulation of materials (i.e., semantic relatedness) was successful. However, it was important to bear in mind that native speakers and learners can behave differently in an experiment.

Table 13

Summary of Mixed Models for Pilot Data – Semantic Priming Task (Native Speaker)

	m0 (null model)		m1 (maximal)	
Fixed effects	estimate [a] (*SE*)	*t (p)*	estimate (*SE*)	*t (p)*
Intercept	-1.91 (0.06)	-34.38 (<.001)	-1.94 (0.06)	-34.91 (<.0001)
Condition			0.07 (0.03)	2.66 (.02)
Random effects	Variance (*SD*)	Intercept-slope correlation	Variance (*SD*)	Intercept-slope correlation
By-participant intercept	0.05 (0.22)		0.05 (0.21)	
By-item intercept	0.003 (0.06)		0.01 (0.09)	
By- participant slope for condition			0.003 (0.05)	.25
By- item slope for condition			0.01 (0.08)	-.93
Residual	0.10 (0.32)		0.10 (0.32)	
AIC	666		653	

Note. [a] all estimates were multiplied by 1000 for easier reading.

Overall, the results from native speakers and learners offered a mixed message in that priming was only observed in the native speaker data. At the same time, I decided to proceed to the main data collection because of, in part, resources available (time and funds) and, in part, a few reasons for optimism. In the main data collection, I had planned to have more than 100 learner participants to the extent that resources allow. This sample size would reduce the sampling errors, providing more meaningful evidence. I had also planned to incorporate random effects associated with both participants and items, putting me in a position to inspect

the variability in eliciting priming between different items. This information would become crucial in assessing what item(s) consistently elicit random responses. Removal of these item could not only improve data quality, but also reduce the noise (random variability) in the data set to allow the genuine effects to emerge (e.g., Siegelman et al., 2017) (see more discussion in Data Analysis).

Despite the optimism, overall priming was still not observed in the main round of data collection with 145 participants even after item screening (see Chapter 3 for details). This result casted doubt on what was being measures by the task. Therefore, I decided to drop this measure from the present study. I will report the results for this task in Chapter 3 and return to discussing this decision in Chapter 5.

Self-reported Proficiency

Participants self-reported their perceived general proficiency levels on a 10-level Linkert scale where 1 was labelled as "Total beginner" and 10 was labelled as "Native-like." The decision to use a self-assessment questionnaire item was largely motivated by the resources available for the present research. Although it was not a formal assessment, the self-reported rating represented a proxy for their language performance. In a recent meta-analysis involving 67 primary studies and 68, 500 participants, Li and Zhang (2021) found an overall, moderate correlation at .47 ($p < .01$) between self-assessment and language performance, confirming the value and validity of self-assessment, especially when resources present a limitation for researchers to administer a formal language proficiency test.

Procedure

I collected data online through Gorilla, as reported above. Interested participants first completed a screening survey where they read information about the study (e.g., procedure, potential risks, payment, and so on) and expressed consent to participate. They also reported their status as (1) a non-native speaker of English and (2) a current student at an American university. When participants fulfilled these two inclusion criteria, I then entered their university email to Gorilla which sent a message to the participants, directing them to log into the system. Before the start of the experiment, they read the study's information and offered consent again. They also provided their demographic information (e.g., age, length of residence in the US).

The whole experiment consisted of a battery of five tests (see Measures above), administered on two separate days with at least 48 hours apart. In determining the order of administration, I considered the extent to which a given task might have an effect on subsequent ones. This was an important consideration given that multiple testing of the same critical words appeared to be "unavoidable" in this line of research (González-Fernández & Schmitt, 2020, p. 23). I also considered both the task requirement (production *vs.* recognition), the proportion of the critical words in the whole stimuli set, and potential impact of fatigue on the data quality and on participant attrition. I decided that the three sensitive tasks generating reaction time data should precede the paper-based tests because the former would be more sensitive to any potential effects of fatigue. Then, the form-meaning productive task should precede the form-meaning receptive task. Between these three psycholinguistic tasks, I counter-balanced the order to cancel out any potential ordering effects. On Day 1, then,

participants took part in the three reaction time-based tasks. Forty-eight hours later, the system sent a reminder email to the participant who was then able to log in and complete the two paper-based tests. At the end of the study, participants entered their payment information (e.g., electric payment account).

Data Analysis

In this section, I detail the data analysis plan. I will first discuss the analyses for specific tasks with the ultimate goal to obtain the most accurate and reliable measures for each task. Then, I describe the overall CFA data analysis to address the research question. Table 14 summarizes the number of data points for each participant in each of the five tasks.

Table 14
Summary of Number of Data Points for Each Participant

Task	No. of Data Points Analyzed	Details
Form-Meaning Receptive Test	40 accuracy data points	One observation for each critical word (K = 40)
Form-Meaning Productive Test	40 accuracy data points	One observation for each critical word (K = 40)
Yes-No RT Test	80 accuracy and 40 reaction time data points	One accuracy and one reaction time data point for each critical word (*k* = 40), and one accuracy data point for each non-word (*k* = 40)
Masked Repetition Priming Task	80 reaction time data points	Two reaction time data points for each critical word (K = 40)
Semantic Priming Task	80 reaction time data points	Two reaction time data points for each critical word (K = 40)

The Form-Meaning Receptive Test

This task afforded accuracy data for all 40 critical words. Coding was conducted automatically through matching the key and the option chosen by the participant. To this end, I used the programming language R to minimize human errors in the process. I adopted the suggested key in the Vocabulary Size Test as the expected answers. I coded responses that matched with the key as 1 and those that did not match as 0. Since all items required a response, there were no missing data. The system also did not allow selection of more than one option for most items (38 out of 40). However, due to a programming error, participants were able to choose multiple answers for two items. In these cases, the participant scored 0 (incorrect) for that item because they did not follow the test instructions, regardless of whether the correct option was chosen or not. This scenario represented 0.07% of the data.

I then conducted a by-participant analysis to screen out participants who might not have paid sufficient attention and/or did not have the proficiency levels to take part in the study meaningfully. In particular, when a participant scored below 25% (chance level given four options), I coded their data for this task as missing. After the by-participant analysis, I inspected the items. I first removed items that had no item variance (i.e., those that all participants responded to (in)correctly) because these items could not discriminate participants' ability. Then, I started by examining the instrument reliability in terms of both Cronbach's alpha and McDonald's omega (e.g., McNeish, 2018; Raykov & Marcoulides, 2019). I also submitted the data to a dichotomous basic Rasch analysis (Rasch, 1960) with an aim to identify unsatisfactory items.

Rasch analysis is a commonly used statistical technique in psychological science (e.g., Müller, 2020) and language testing (e.g., Aryadoust et al., 2020; McNamara & Knoch, 2012) particularly to evaluate the psychometric properties of assessment items. A Rasch model predicts the probability of a learner answering an item correctly. To do so, it estimates parameters for both learner ability and item difficulty on the same scale. When the difficulty level of an item coincides with a person's ability level, the person has a .50 chance of answering it correctly. When one has a higher ability in the construct that the test is set out to measure, one performs better on the test because one has a high probability of correctly answering more items across levels of difficulty (see Aryadoust et al., 2020 for an overview). At first sight, this might seem intuitive, but a couple of statistical assumptions are often made: First, all items are assumed to measure one single construct. This assumption of unidimensionality has sometimes been regarded as "too stringent" for language assessments because constructs, such as reading and listening, are fundamentally multidimensional (Aryadoust et al., 2020, p. 4). Another assumption that might not be easily met is local independence, meaning that all items measuring a unidimensional latent trait (i.e., person ability) are correlated with each other only because of the trait. Once the trait is controlled for, there should no longer be correlation left between these items. In statistical terms, the residuals (errors) associated with the items when regressed on learner ability should not correlate with those of other items (e.g., Aryadoust et al., 2020).

Despite some rather strong assumptions, a Rasch model provides useful information about the learners and the items. First, by taking into account item difficulty, estimates of learner ability can be more accurate than using a sum score that ignores item characteristics.

Second, a Rasch model provides statistics for evaluation of the measurement. For example, infit and outfit metrics helps researchers identify items that elicit erratic responses both near or far from the learner ability (i.e., on- and off-target responses). In particular, the mean square index, summarizing the standardized residuals based on the estimates of response probabilities, can offer insight into the amount of noise (random variability) in the data for each item. In particular, when the mean square value is too low, this indicates that an item is overfitted (too little error), potentially signaling the redundancy of the item (Wright & Linacre, 1994). In contrast, an underfitted item (too much error) reveals itself by having a high mean square value, potentially due to random lucky guesses (Wright & Linacre, 1994).

Using basic Rasch models, I inspected these fit statistics for each item. There are two ways to evaluate item fit: One can use rule-of-thumb critical values, or one can conduct a formal test and compare the resulting fit statistics against a normal distribution (Müller, 2020). Although any rule-of-thumbs are almost always controversial among statisticians, I adopted this former approach because it is more straightforward for applied researchers. Also, the exact distribution of the fit statistics is not yet very well known; therefore, conclusions from a formal test may or may not be appropriate (Müller, 2020). When determining the appropriate critical values for the present context, I first considered the scale of these item fit statistics. Both infit and outfit statistics are on a scale between 0 to positive infinity, with 1 being the ideal value indicating no or little distortion in the measurement system (Wright & Linacre, 1994). A value higher than 1 indicates underfitted items, while a value below 1 signals overfitting (redundancy) (Wright & Linacre, 1994). Wright and Linacre (1994) suggested that researchers should first focus on outfit (off-target responses) before infit (on-target responses) statistics, and on high

values (underfitting) before low (overfitting) values. In terms of an appropriate range, 0.5 to 1.5 is most commonly used in language testing contexts (Aryadoust et al., 2020). But, the authors also recommended 0.7 as the lower bound for low-stakes dichotomous tests. For the upper bound, the authors suggested following the equation provided by Smith et al. (1998) which takes the total number of items into consideration. In the case of 40 items, for example, the threshold can be set at 1.95 ($1 + 6 / \sqrt{40} = 1.95$). I kept these guidelines in mind when inspecting the items, while also considering the test reliability measures. I then excluded items that were considered not psychometrically satisfactory. I then repeated the analysis and saved the learner ability parameters from the refitted Rasch model for further CFA analysis.

The Form-Meaning Productive Test

This test afforded 40 accuracy data points for each participant. I coded each correct answer as 1 and incorrect answers as 0. I followed Laufer and Nation (1999) in ignoring "[m]inor spelling… and grammatical mistakes" (p. 38 – 39). In terms of grammatical mistakes, I only accepted mistakes in inflections (e.g., plural or tenses) and capitalization. Incorrect parts of speech (i.e., derivational errors) were marked as incorrect. From all responses, I first summarized the unique responses for all items with an aim to create a coding scheme that specifies acceptable alternative responses. Due to the subjectivity involved, I invited a second rater, who is an experienced teacher of English as a second language, to discuss and co-construct the coding scheme with me. Then, I used the programming language R to match participants' response with this coding scheme. I followed the same analytic procedure as described above. In short, I first conducted a by-participant analysis, followed by an assessment of test reliability. Then, I inspected item characteristics with a Rash analysis. I saved the learner

parameters from the Rasch model built upon a final data set after any item exclusion for further analysis.

Yes-No RT Test

This test provided both accuracy and reaction time data automatically logged by Gorilla. I first inspected the false alarm (incorrect responses to non-words) rates by participants. Any participants who had a false alarm rate higher than 0.50 were deemed as either not paying sufficient attention and/or did not have the proficiency levels to complete the study meaningfully. I coded their data as missing for the CFA. In terms of item quality, I followed X. Zhang et al. (2020) in fitting two separate Rasch models to the real and non-word data. Unsatisfactory items were removed for further analysis. I also computed reliability estimates for both data sets. The inspection procedure was the same as described above.

In analyzing the accuracy data, I consulted the literature on the scoring of the traditional, paper-based Yes-No test (e.g., Huibregtse et al., 2002; X. Zhang et al., 2020). I first computed the hit (correct responses to real words) and false alarm (incorrect responses to non-words) rates for each individual. As alluded to in the Measures section, it was important to take guessing into account when computing an index to reflect one's ability. I used individuals' false alarm rate (guesses on non-words) as an operational measure of guessing on real words (X. Zhang et al., 2020; cf. Stubbe, 2012). In the literature on scoring a paper-based Yes-No test, different formulas have been proposed to calculate a measure to index ability. These formulas include the simple hits-minus-false-alarms rule, the correction for guessing formula, the delta m, and the index of signal detection (see a review in Huibregtse et al., 2002). On the one hand, there appeared little consensus as to the best-performing scoring method (Pellicer-Sánchez &

Schmitt, 2012), and little meaningful differences were found in Mochida and Harrington (2006). On the other, a recent study by X. Zhang et al. (2020) reported a range of correlations between Yes-No Test scores adjusted by these different formulas and two reference tests: an MC Vocabulary Size Test and a translation task. The overall corrections ranged from .289 (when using delta m) to .621 (when using the index of signal detection). An additional consideration in their study was the false alarm rates observed in the sample. The authors computed correlations for those one with high and low false alarm rates. As expected, the group with a high false alarm rate showed much weaker correlations (.297 - .530) than the group with a low false alarm rate (.636 - .708). This was because participants in the high false alarm group were guessing randomly more, hence there was more noise in their data, weakening the correlations. In both group and overall analyses, the index of signal detection had the strongest correlations with the vocabulary size test, which was essentially the form-meaning receptive test in the present study. For this reason, I chose this formula to compute individual's performance for further CFA analysis:

$$I_{SDT} = 1 - \frac{4(1-f) \; - \; 2(h-f)\,(1+h-f)}{4(1-f) \; - \; (h-f)\,(1+h-f)}$$

where *h* is the hit rate and *f* is the false alarm rate.

In analyzing the reaction time data, I included only the correct, real word trials, following standard practice in psycholinguistic research because recognition of word and non-word involves different processes. I further trimmed the spuriously short (< 300 ms) and long (>

2500 ms) reaction times (Jiang, 2013) because they do not reflect the lexical processing under investigation. I then computed a mean reaction time for each participant for further analysis.

Masked Repetition Priming

Although the priming effects were the primary focus of this task, I first inspected the false alarm rate of individual participants to exclude anyone who guessed on more than 50% of the non-words. With regards to the priming effects, the present task has mostly been used in experimental contexts where a group-level effect is examined (e.g., Draheim et al., 2019; Rouder & Haaf, 2019; Siegelman et al., 2017). Psychometric properties of experimental tasks, such as this one, have recently raised concerns in terms of their level of reliability (Draheim et al., 2019). In more general terms, Siegelman et al. (2017) discussed three design features of experimental tasks that make them perform somewhat unreliably: first, there are often a small number of trials, leading to little room for between-participant variance, which is an important analytical component in individual differences research. Second, participants oftentimes perform at chance levels on some, if not most, items. When that is the case, any expected effects, especially those with small effect sizes, can be buried in the random variability (noise) of the data. Finally, most of the items often have similar levels of difficulty, which limits the ability of the tasks to differentiate participants' ability.

More directly relevant to the present task is the discussion of the same issue in reaction time-based research in behavioral science by Draheim et al. (2019). In general, researchers often use a difference in reaction time to demonstrate an effect. In the present task, for example, participants were expected to respond faster when the target was preceded by a related, repetition prime than when preceded by an unrelated prime. Then, the difference in

reaction time could be computed as that in the unrelated condition minus that in the related condition, whereby I expected a positive net value (i.e., faster responses in the related trial). Draheim et al. (2019) pointed out that the primary issue with using a difference reaction time is that "as the *correlation* between the two component scores *increases* [reaction times in each condition], the reliability of the resulting difference score *decreases*" (p. 511, emphasis original). It is because subtraction removes systematic variance due to the participant's characteristics, hence increasing the proportion of error variance in the data (Draheim et al., 2019). For these reasons, the authors concluded that "difference scores are poorly suited for this purpose [of individual differences research]" (Draheim et al., 2019, p. 513).

To mitigate the potential issue at hand (i.e., a potential low reliability), I adopted three strategies: first, I minimized random variability in the data through an item analysis. Second, I engaged in mixed-effects modeling at the trial level (Rouder & Haaf, 2019). Finally, when constructing the mixed models, I engaged in model criticism to maximize model fit (Baayen & Milin, 2010) as a means to further reduce random variability in the data.

As pointed out by Siegelman et al. (2017), individual items might elicit random performance from participants. Although this task had been piloted and the results were encouraging, the reliability of results might still benefit from removal of unsatisfactory items (reducing noise in the data). However, there have not been explicit guidelines that psycholinguists rely on in terms of item inspection, in part because researchers are often interested in group-level effects. Therefore, to inspect item quality, I drew on the recommendations of using mixed-effects modeling in analyzing these data by Rouder and Haaf (2019).

Mixed-effects modeling is a multi-equation technique which has the ability to model and account for dependency between observations due to a nested data structure (e.g., Gelman & Hill, 2007; Hox et al., 2018). A nested data set is one where lower-level observations correlate with each other because they are associated with a higher-level unit. The typical textbook example in educational contexts is one where students (level-1 units) are nested within classes (level-2 units). In psycholinguistic research, one participant contributes multiple data points to the data set (as they respond to multiple items); at the same time, each item elicits multiple responses from the sample of participants. In that case, the data can be said to be cross-nested, and random effects by participants and by items should be incorporated simultaneously in the data analysis in order to account for the dependency (e.g., Baayen et al., 2008). Conceptually, the mixed model estimates an intercept and a slope value (as well as the correlation between them) for each participant and item (level-2 units). The intercept value represents the mean (transformed) reaction time for the participant or item in the reference condition (i.e., the related condition for the present study). The slope value represented the simple effect of condition (i.e., the difference in reaction time between the related and unrelation conditions) for a particular participant or item as deviated from the fixed effects estimate.

Rouder and Haaf (2019) demonstrated that estimates associated with participants tend to be more accurate and reliable when the item-related variability is incorporated in the model as random effects. Put differently, after accounting for item characteristics, researchers can more accurately, reliably model an effect at an individual's level. To do that, one can fit a maximal model where all relevant random intercepts and slopes (both by participants and by items) are entered and allowed to vary (e.g., Barr et al., 2013). This analytical approach also

lends itself as tool for a model-based item inspection. In the case of a maximal model, the random slopes for condition by items represent the effects of condition for a specific item. Then, one can inspect the random slopes by items to identify (and potentially remove) which items elicit random responses from the sample of participants. However, excessive item removal could result in an undesired decrease in reliability. Therefore, caution was exercised in balancing identifying and removing potential noise in the data and retaining as much statistical information as possible.

In addition, Baayen and Milin (2010) recommended that researchers analyzing reaction time data should engage in model criticism (see Godfroid [2020a] for a similar recommendation for analyzing reading time data obtained from eye tracking). Model criticism is a strategy to treat outliers in reaction time data. Researchers first fit a mixed model to the data. From this initial model, researchers remove observations with "an absolute standardized residual exceeding 2.5 standard deviations" (Baayen & Milin, 2010, p. 17). This strategy is compared favorably with traditional outlier removal procedures (e.g., removing data points that are beyond 2.5 *SD* of the mean before a model is fitted). The authors demonstrated the advantage of model criticism in terms of the number of observations needed to be removed (or ability to retain as much information as possible) and model fit (i.e., R^2 values). The key rationale behind this strategy is that removal of outliers is more principled when informed by a model, taking all relevant conditional effects into account in a parsimonious manner. On this account, this model-based analytic procedure has the potential to address the issue of reliability associated with experimental tasks. This is because a better model fit (as a result of engaging in model criticism) means less random variability in the data, potentially leading to more accurate

estimates of the parameters associated with the participants, which represent the individual differences measure for the present task. Taken together, the ultimate goal here was to minimize random errors in the model from which I saved the calibrated participant-related estimates as a measure for this task. This was achieved by an item inspection procedure and use of mixed-effects models for which I engaged in model-based outlier removal.

In analyzing the data, I only included correct responses and the trimmed reaction times using the same thresholds as other reaction time tasks (i.e., 300 ms < RT < 2500 ms). Then, I constructed mixed effects models according to the modeling procedure reported in the Measures section. Then, I engaged in model criticism and refitted the model with outliers removed. From the refitted model, I inspected the random slopes for each item. I also split the data into two halves and repeated the same procedure to obtain split-half reliability for the by-participant random slopes. In an iterative manner, I assessed the impact of item removal on the split-half reliability and decided on the final data set (i.e., to remove the item in question or not). I returned to the mixed model constructed from this final data set and saved the by-participants slopes for further CFA analysis.

Semantic Priming

The data analysis procedure was identical to that for the masked repetition priming.

Main CFAs and SEMs

Using the results of the individual tasks, I engaged in a confirmatory factor analysis to address the first research questions (RQ1a and RQ1b). Confirmatory factor analysis is often contrasted with exploratory factor analysis, both of which can be regarded as a member of a

family of techniques known as factor analysis (e.g., Loewen & Gönülal, 2015). The main use of factor analysis is to lay bare the underlying relationships between a set of (observed) variables. These relationships are believed to be driven by a more parsimonious number of latent (unobserved) variables, known as factors (e.g., Brown, 2015). By specifying the factor structure (e.g., a one- *vs.* two-factor solution) as informed by prior evidence and theory, researchers are in a position to test the psychometric dimensionality of a set of measures, lending itself an appropriate tool for construct validation research (R. Ellis & Loewen, 2007; Isemonger, 2007; Vafaee & Kachinske, 2019). In addition, researchers can take advantage of the modeling flexibility in CFA to evaluate construct validity in the light of different assessment methods by introducing (and partially out) method effects in the model (Brown, 2015).

In terms of model specification, I initially had a total of six measures: (1) the form-meaning receptive test (Recep), (2) the form-mean productive test (Prod), (3) the Yes-No RT test – accuracy data (Yes-No-Acc), (4) the Yes-No RT test – reaction time data (Yes-No-RT), (5) the masked repetition priming task (RepPrim), and (6) the semantic priming task (SemPrim). As reported, I dropped the semantic priming task (see results in Chapter 3 and discussion in Chapter 5) because overall priming was not observed. Therefore, I had a total of five indicators in the model. In the variance-covariance matrix, then, there were 15 pieces of information (five variances on the diagonal and 10 covariances [4+3+2+1] between the indicators off the diagonal). For identification purposes, the CFA models can have at most 14 freely estimated parameters so that the model is over-identified for model fit evaluation (i.e., assessing the extent to which the model is an acceptable representation of the data).

In order to address RQ1a, the extent to which there were two separate dimensions of lexical knowledge based on awareness (explicit *vs.* implicit), I had had an initial plan to fit a two-factor model (one factor labelled as explicit and another labelled as implicit). In this model, the four explicit measures (Recep, Prod, Yes-No-Acc and Yes-No-RT) would load onto the explicit factor, while the two priming tasks would load onto the implicit factor. This two-factor model would be compared against the rivalry one-factor model where all six measures loaded onto one single vocabulary factor. However, the semantic priming task was dropped as reported in the Measures section. This decision led to identification issues with the implicit factor because it had then one indicator remaining. As a result, two models were tested to shed light on the research question, but they were limited in terms of offering conclusive evidence. In model CFA-M1, all four explicit measures loaded a single latent variable in a one-factor model. I labelled the factor as vocabulary. Then, in CFA-M2, I added repetition priming to examine potential evidence of uni-dimensionality (i.e., to what extent this implicit measure can be placed on the same psychometric dimension as the other explicit tasks). In this model, I allowed the residuals of the mean RT and the repetition priming to correlate because both measures were reaction time-based. This additional parameter represented a multitrait–multimethod approach to partial out any method effects (e.g., Brown, 2015). Although a direct model comparison between CFA-M1 and CFA-M2 would be inappropriate (because the two models had different sets of indicators), the model fit of these two models could still provide some useful information. Specifically, if CFA-M2 could not produce a good fit, there would then be evidence against a uni-dimensionality view. To address RQ1b, I fitted a two-factor model (CFA-M3) where the untimed tasks (Recep and Prod) loaded onto a knowledge factor and the timed

measures (Yes-No-Acc, Yes-No-RT, and RepPrim) loaded onto a strength factor. Again, methods effects were accounted for by allowing the residual terms of the two reaction time-based measures to correlate. The rivalry model to this one was the one-factor CFA-M2 described above. I summarize the hypothesized CFA models in Table 15.

Table 15
Summary of Hypothesized CFA Models

Explicit *vs.* implicit				
Model	Vocabulary			Method Effects
CFA-M1 (one factor)	Recep; Prod; Yes-No-Acc; Yes-No-RT			
CFA-M2 (one factor)	Recep; Prod; Yes-No-Acc; Yes-No-RT; RepPrim			Yes-No-RT ~~ RepPrim
Model	Vocabulary	Knowledge	Strength	Method Effects
CFA-M2 (one factor)	Recep; Prod; Yes-No-Acc; Yes-No-RT; RepPrim			Yes-No-RT ~~ RepPrim
CFA-M3 (two factors)		Recep; Prod	Yes-No-Acc; Yes-No-RT; RepPrim	Yes-No-RT ~~ RepPrim

In terms of RQ2, concerning the predictive validity of the different vocabulary constructs, I used a full structural equation modelling (SEM) approach where I regressed self-reported proficiency on the CFA (measurement) models described above. Specifically, in SEM-M1a, I used CFA-M1 to predict self-reported proficiency. I included the repetition priming measure as an observed variable (an additional predictor) in SEM-M1b. As the next step, I fitted SEM-M2 where I used CFA-M2 to predict self-reported proficiency. With regards to the

knowledge *vs.* strength distinction, I fitted SEM-M3 where the two factors in CFA-M3 predicted self-reported proficiency. Finally, all five measures were used as observed variables to predict proficiency in SEM-M4 (i.e., the same as a multiple regression model where the outcome was proficiency, and the five measures were predictors).

In estimating all CFA and SEM models, I used case-wise (full information) maximum likelihood estimation with robust (Huber-White) standard errors correction. The case-wise estimation allows a participant's data to contribute to the model even when they have missing data for a specific task, and the robust standard errors mitigate the impact on non-normality in the data (e.g., Brown, 2015; Kline, 2015).

To address the first research questions, I relied on evidence of good model fit. I examined both global and local fit (for the CFA models). For global fit, I used the following fit statistics: model chi-square with its degree of freedom and associated *p* value (> .05), Root Mean Square Error of Approximation (RMSEA) and its confidence intervals (< .05), Comparative Fit Index (CFI) (> .95), and Standardized Root Mean Square Residual (SRMR) (< .08) (e.g., Hu & Bentler, 1999). In addition to global indices, I also inspected the factor loadings, the standardized residuals (< |1.96|), and the modification indices (< 3.84) to assess potential local misfit. For the second research questions, I inspected the R^2 value for the self-reported proficiency measures. These values represented the explanatory power of a given conceptualization of the vocabulary construct in accounting for the outcome (i.e., proficiency). For illustration, I visualize the CFA-M2 and SEM-M3 Figure 2 and Figure 3, respectively.

Figure 2

Visualization of CFA-M2

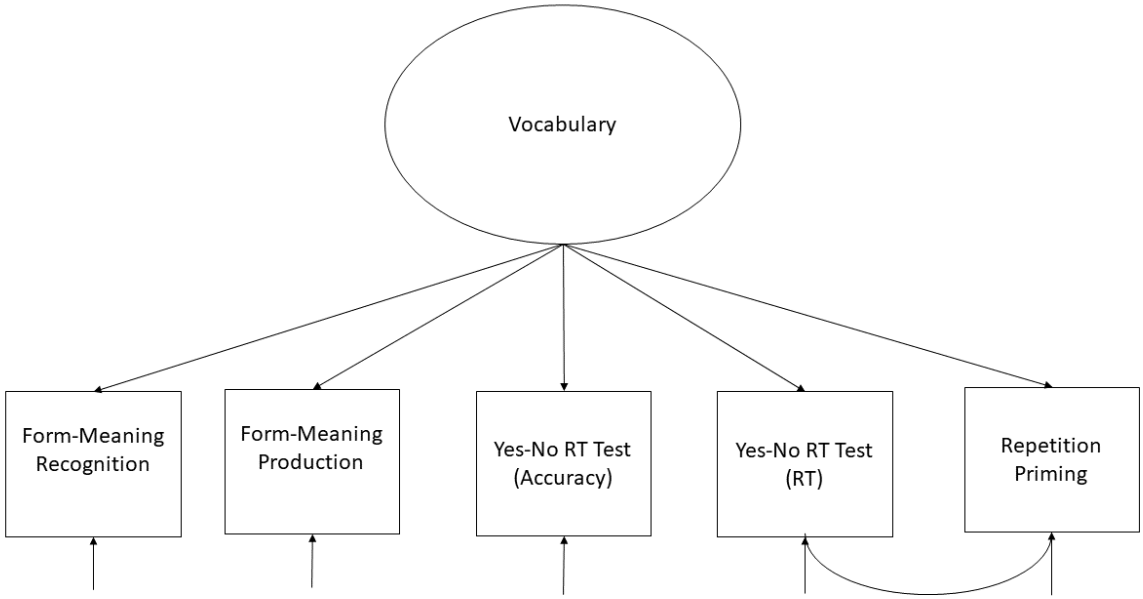
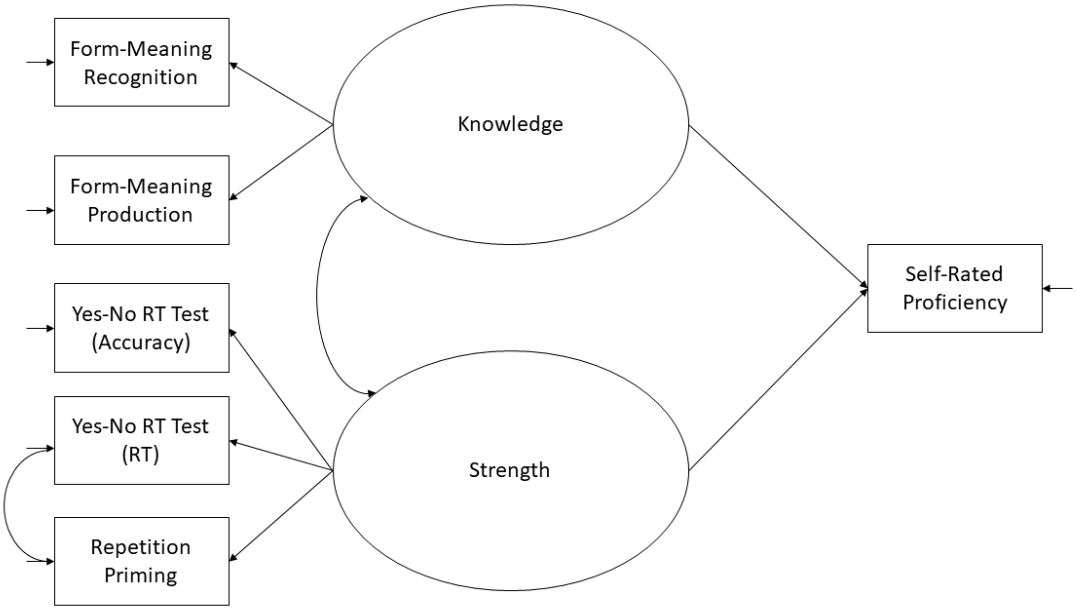


Figure 3

Visualization of SEM-M3



Analysis Software Packages

In Table 16, I list all software packages used in data wrangling and analysis:

Table 16

Summary of Software Packages Used

Tasks	Package Name (Version)	Reference
General use	R (4.0.2)	R Core Team (2020)
General use	RStudio (1.3.1093)	RStudio Team (2020)
Data wrangling	tidyverse (1.30)	Wickham et al. (2019)
Data visualization	ggplot2 (3.3.3)	Wickham (2016)
Summary and reliability	psych (2.0.12)	Revelle (2020)
Mixed-effects modeling	lme4 (1.1-26)	Bates et al. (2015)
	lmerTest (3.1-3)	Kuznetsova et al. (2017)
	performance (0.7.0)	Lüdecke et al. (2020)
	brms (2.14.4)	(Bürkner, 2018)
Rasch analysis	eRm (1.0-2)	Mair and Hatzinger (2007)
CFA	lavaan (0.6-7)	Rosseel (2012)

CHAPTER 3: RESULTS (INDIVIDUAL TASKS)

In this chapter, I report results of the data analysis of the five individual tasks. These results also represent the data preparation process towards the main CFA and SEM which will be covered in the next chapter. As mentioned, the goal of these analyses was to maximize the accuracy and reliability of the measure indexing participants' performance on each test.

The Form-Meaning Receptive Test

Based on the by-participant analysis, one participant was excluded based on below-chance performance (25%). However, the analysis suggested a ceiling effect, creating a left-skewed distribution (see upper panel of Figure 4). I present the descriptive statistics in Table 17.

Table 17
Descriptives for the Form-Meaning Receptive Test

N = 144	Mean (*SD*) [95% CIs]	Range
Total Score (*K* = 40)	37.27 (3.95) [36.62, 37.92]	15 – 40
Total Score (*K* = 35)	32.34 (3.81) [31.71, 32.97]	11 – 35

I first inspected the item variances before computing reliability measures. All items had some variance. With this full 40-item data set, Cronbach's alpha was .90, and McDonald's omega was .91. At first sight, these figures suggested very good reliability, but caution was exercised because of the ceiling effect observed. In other words, many items can potentially be redundant, causing an undesired inflation of the reliability (Wright & Linacre, 1994). With this in mind, I proceeded to the dichotomous basic Rasch analysis. I present the person-item map in

Figure 4. The upper panel represents the left-skewed distribution of the person ability parameters, indicating the ceiling effect. In the lower panel, item difficulty is visualized in relation to the amount of person ability required to answer the item correctly (x-axis). Therefore, items whose data point is plotted left to the graph are relatively easy items (or items that require less person ability to answer correctly). The difficult items have their data point towards the right-hand side of the graph. Due to the ceiling effect, most test takers had an ability higher than the item difficulty of most items. At the same time, there was still a range of item difficulty, with some potentially redundant items (being too easy) clustering in the top right corner of the panel.

In terms of the item fit statistics, given 40 items in this data set, the upper threshold was 1.95 (1+6/$\sqrt{40} = 1.95$), following Smith et al. (1998). Inspecting the outfit statistics first, following recommendations by Wright and Linacre (1994), one item (*stone*) was underfitted, with its outfit statistic exceeding the threshold of 1.95 (outfit = 2.67), potentially indicating this item was often guessed correctly. In terms of the lower bound of the statistics (signaling overfitting [or redundancy]), I first adopted the commonly used threshold of 0.5 (Aryadoust et al., 2020). Four items were below 0.10. They corresponded to the items clustering in the top right corner of the person-item map in Figure 4. Another 13 items had values that ranged from 0.11 to 0.49. As mentioned, these items with a low item fit statistic represented redundant items. However, they did not distort the measurement system (Wright & Linacre, 1994). I also eye-balled the infit statistics, which showed no additional issues.

Since item removal generally results in a lower test reliability, I examined the impact of removing different number of items. I first removed *stone* based on a high outfit value and the

four items with overly low outfit values. In total, then, five items were removed, leaving 35 items on the test. With this data set, Cronbach's alpha for this subset was .89, and McDonald's omega was .90. I fitted the Rasch model to this 35-item data set again. Outfit values ranged from 0.11 to 1.92. Twelve items fell outside the critical threshold range of 0.5 to 2.01, but only on the lower end). I then assessed the extent to which further removal of these items would be beneficial. Reliability estimates for the 23-item test dived to .79 for both Cronbach's alpha and McDonald's omega. In addition, the total proportion of variance explained by the Rasch model dipped slightly to 37.46% (down from 39.94% in the 35-item test). Although a reliability of .79 in the 23-item test can still be deemed as acceptable, I decided to use the data set with 35 items because it had a better Rasch model fit. I present the person-item map for the Rasch analysis with this data set in Figure 5. At the same time, I acknowledge that some items with relatively low outfit statistics could inflate the reliability of the 35-item test. I saved the person ability parameters, which correlated with the raw score at .90, for further analysis.

The Form-Meaning Productive Test

In total, there were 1086 unique responses for all 40 items. To construct a coding scheme, a second rater and I assessed the acceptability of each of these responses separately with reference to the general rule (i.e., only minor spelling and inflectional errors were to be accepted). The initial agreement rate was 97.4%. Cohen's Kappa (interrater) reliability was .91 after accounting for chance agreement. We revisited all instances of disagreement. For all but four disagreements, we were able to self-correct the coding and reach an agreement without a discussion. The four occasions that required a discussion involved participants putting in more than one word and mistakes concerned with the -ed *vs.* -ing forms (e.g., paved *vs.* paving). We

resolved these issues, and the resulting coding scheme was implemented in the scoring of the test using the programming language R.

Table 18
Descriptives for the Form-Meaning Productive Test

N = 138	Mean (*SD*) [95% CIs]	Range
Total Score (*K* = 40)	29.05 (6.95) [27.9, 30.2]	11 – 40
Total Score (*K* = 38)	27.19 (6.92) [26.02, 28.35]	9 – 38

The by-participant analysis suggested that seven participants scored very low on the test (raw scores = 0 and 8 [out of 40 items]). I excluded them because their attention and/or proficiency levels were questionable. Then, I generated descriptives as presented in Table 18. All items had some item variance for the computation of reliability measures. Cronbach's alpha and McDonald's omega were both at .88.

From the Rasch analysis, two items were underfitted (*upset* and *cap*) having an outfit statistic exceeding the upper threshold (given 40 items, 1+ 6/ $\sqrt{40}$ = 1.95) (Smith et al., 1998). Two items (*microphone*, and *crab*) were identified as redundant with outfit statistics ranging from 0.38 – 0.45. As with the other tasks, I only removed the underfitted item (k = 2), resulting in a 38-item data set. I then repeated the accuracy analysis again. Cronbach's alpha did not change, while McDonald's omega slightly increased to .89. Item fit issues were not found from the refitted Rasch model (outfit statistics ranging from 0.37 to 1.61 with the model accounting

for 75% of the variance). I saved the person ability parameters for further analysis. I present the person-item map in Figure 6.

Figure 4

Person-Item Map for the Receptive Test (40-Item)

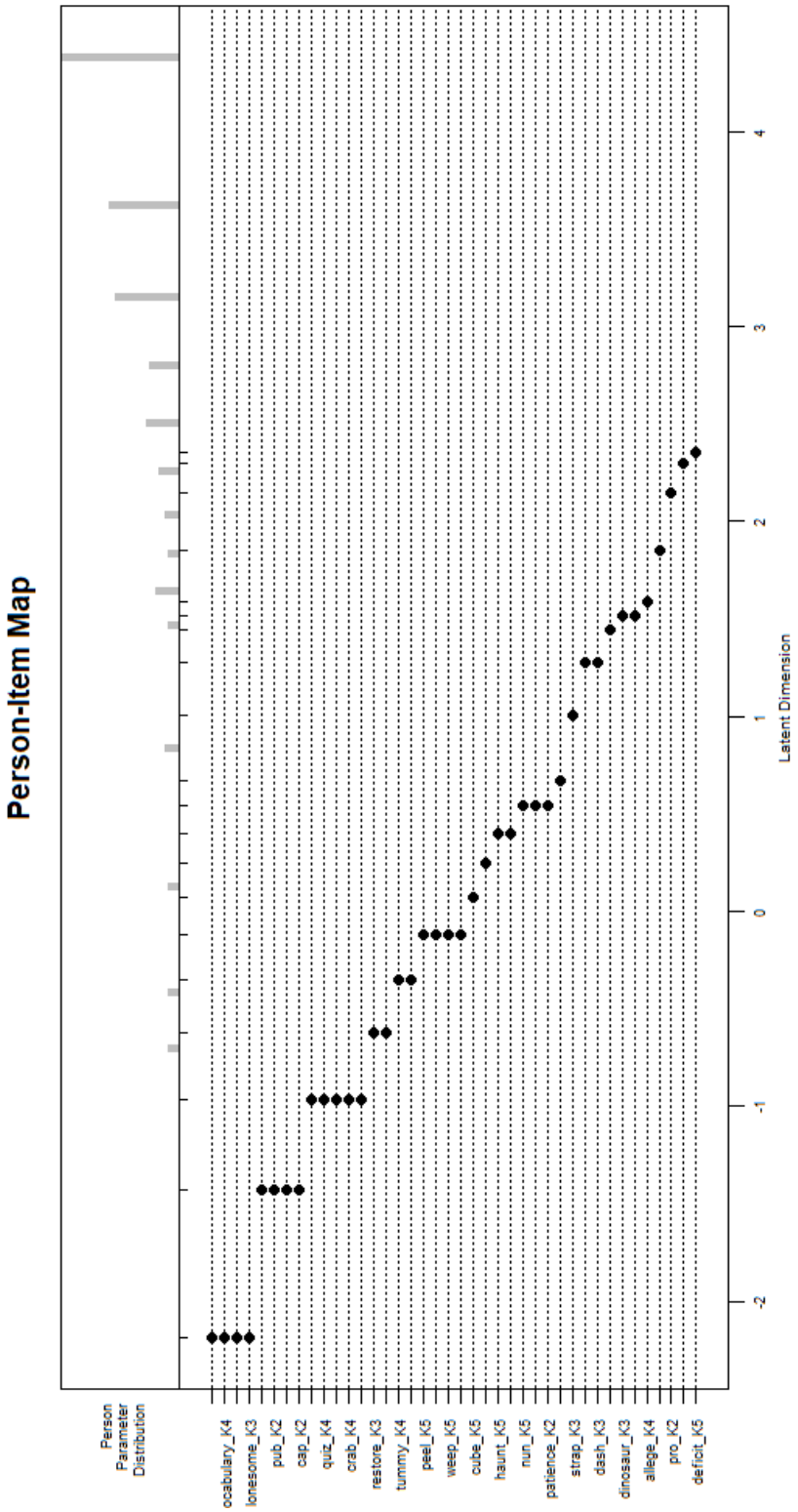


Figure 5

Person-Item Map for the Receptive Test (35-Item)

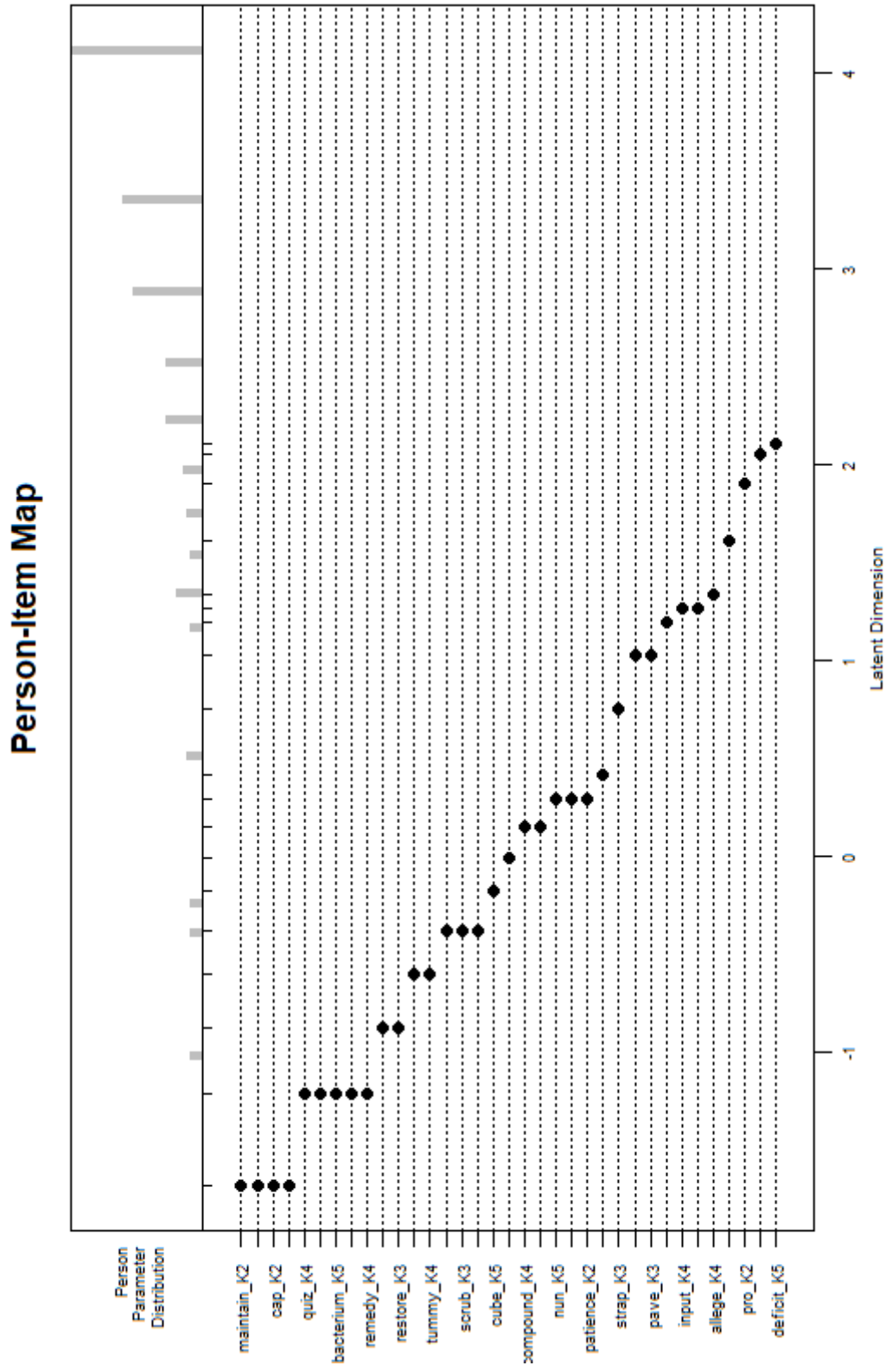
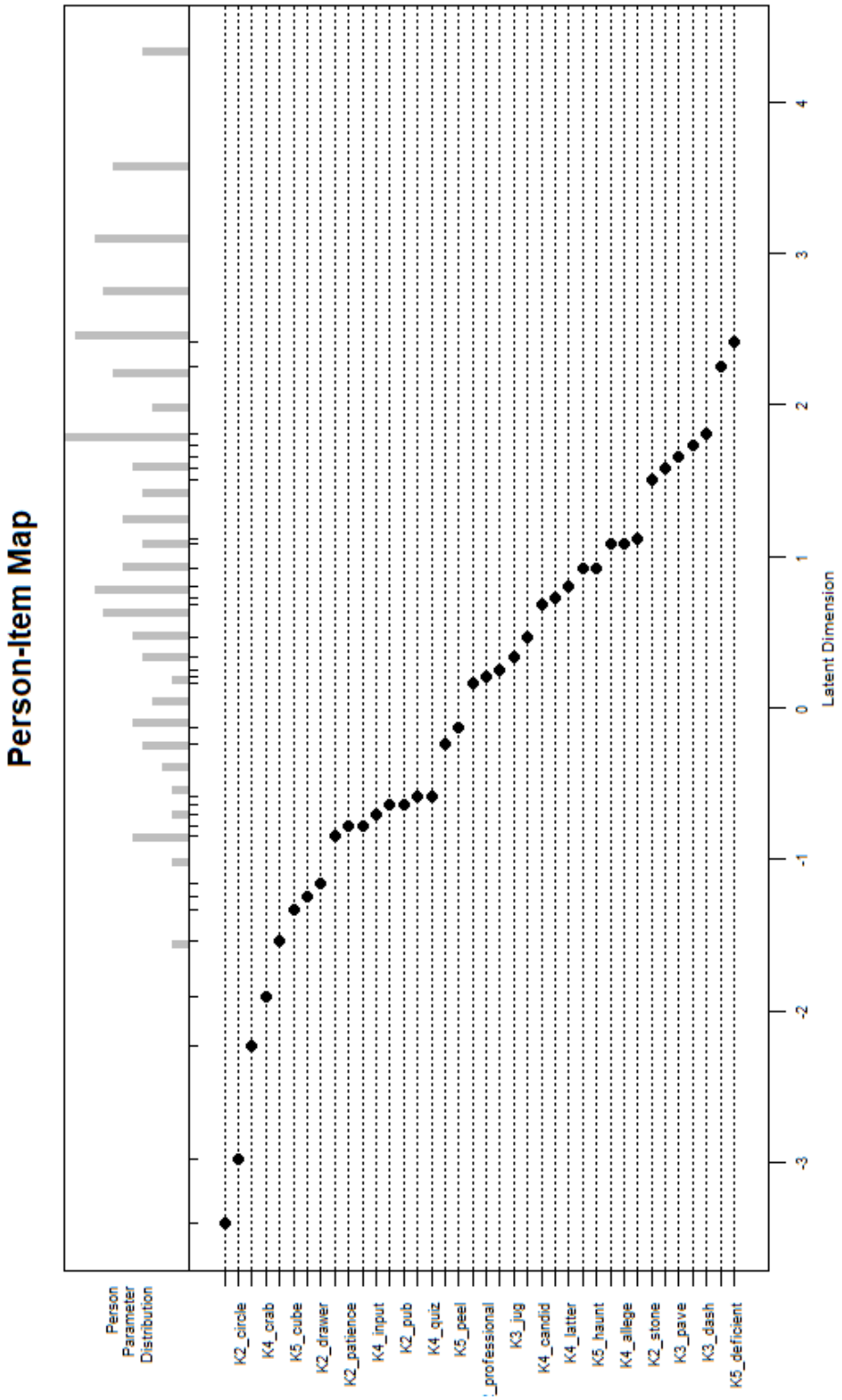


Figure 6

Person-Item Map for the Productive Test (38-Item)



Yes-No RT Test

Accuracy Data

Based on the by-participant analysis, seven participants (out of 145) were excluded based a false alarm rate larger than 0.50, indicating a high level of guessing on the non-words (false alarm rates ranged from 0.58 – 0.90). As with the form-meaning tests, the by-participant analysis suggested a ceiling effect, creating a left-skewed distribution in both the word and non-word data. I present the descriptive statistics in Table 19.

Table 19
Descriptives for the Yes-No RT Test (Accuracy Data)

N = 138	Mean (*SD*) [95% CIs]	Range
Total Score (Word, *K* = 38)	34.46 (4.31) [33.73, 35.18]	16 – 38
Total Score (Word, *K* = 33)	29.65 (4.15) [28.95, 30.35]	13 – 33
Total Score (Non-word, *K* = 40)	36.24 (4.41) [35.50, 36.98]	20 – 40
Total Score (Non-word, *K* = 37)	33.30 (4.29) [32.58, 34.03]	17 – 37

Before computing the reliability measures, I removed two items from the word data (*input,* and *vocabulary*). They had no item variance in that all participants responded to them correctly. No non-words needed removal on the basis of a lack of item variance. With the word and non-word data, Cronbach's alpha and McDonald's omega estimates were all at .86.

In the first round of Rasch analysis with the word data, three items were underfitted (*restore*, *drawer*, and *maintain*), having outfit statistics at 2.09, 2.16, and 4.90, exceeding the upper threshold (given 38 items, 1+ 6/ $\sqrt{38}$ = 1.97) (Smith et al., 1998). One items (*stone*) was

overly redundant with the outfit statistic far under the 0.5 threshold at 0.09. It represented an item that was too easy to participants (see also the person-item map in Figure 7). An additional two items (*cap*, and *peel*) were also identified as redundant with outfit statistics ranging from 0.28 – 0.48. As with the other tasks, I only removed the underfitted (k = 3) and overly redundant (k = 1) items, resulting in a 34-item data set for real words. I then repeated the accuracy analysis again.

With these 34 items, both Cronbach's alpha and McDonald's omega were at .86 (same as the last round). The Rasch model further suggested that *dash* as underfitted (outfit statistics at 2.25 which was larger than the threshold of 2.03 [given 34 items]). Removal of it resulted in a slight dip in Cronbach's alpha to .85, and McDonald's omega was the same at .86. Therefore, I settled on this 33-item data set. The Rasch model constructed with this data set showed acceptable item fit with outfit statistics ranging from 0.33 – 1.81. The total proportion of variance explained by the Rasch model was 51%. I then used this 33-item data set to compute a hit rate (correct responses to words), which was then a basis for the index of signal detection to be submitted to further analysis.

In terms of the non-word data, no item needed removal due to a lack of variance. The initial reliability estimates were both at .86. Rasch analysis indicated that one item (*skoign*) was underfitted (outfit = 2.60). Seven items may be considered as redundant (outfit ranging from 0.26 – 0.49). Since the outfit statistics were not overly low, I arbitrarily decided to use a 0.3 cut-off to balance retaining as many items as possible and reducing noise in the data set. As a result, two items (*prarns,* and *zolved*) were additionally removed. After removal, the two reliability measures dipped slightly to .85 (from .86 with the full set of 40 words). I refitted the

Rasch model, and all items had acceptable fit (outfits: 0.35 – 1.57). The total proportion of variance explained by the Rasch model was 51%. I then used this 37-item data set to compute a false alarm rate (incorrect responses to non-words), which was then the basis for the index of signal detection to be submitted to further analysis.

Reaction Time Data

The full word data set (*K* = 40) was used for this analysis. Before the data analysis, seven participants who had a false alarm rate larger than 0.50 were excluded as discussed above. Filtering correct responses resulted in a loss of 489 observations (9%), echoing the accuracy data reported above. Trimming of reaction times (i.e., 300 ms < RT < 2500 ms) removed 149 observations (3%). After this procedure, one participant had no data left. In Table 20, I present the descriptive statistics for the reaction times. I further divided the data set by two (evenly across frequency bands) to compute a split half reliability. The correlation estimates for the mean raw reaction times were at .87. I then recombined the data set and submitted the overall mean reaction times for further analysis.

Table 20
Descriptive Statistics for the Yes-No Test (Reaction Time Data)

N = 137, *K* = 40	mean (*SD*) [95% CIs]
Reaction Time (ms)	755 (309) [746, 763]

Figure 7
Person-Item Map for the Yes-No RT Test - Word Data (33-item)

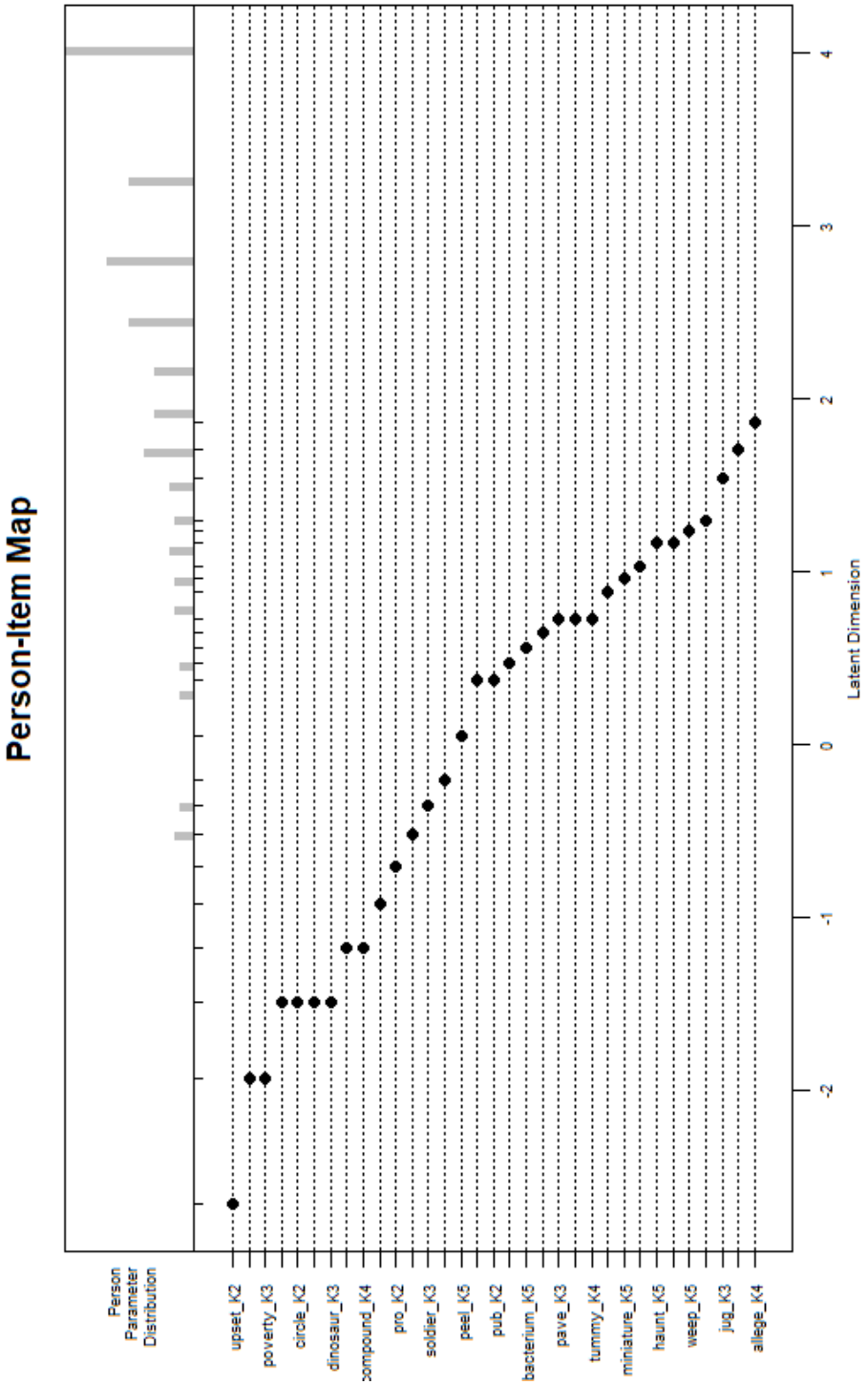
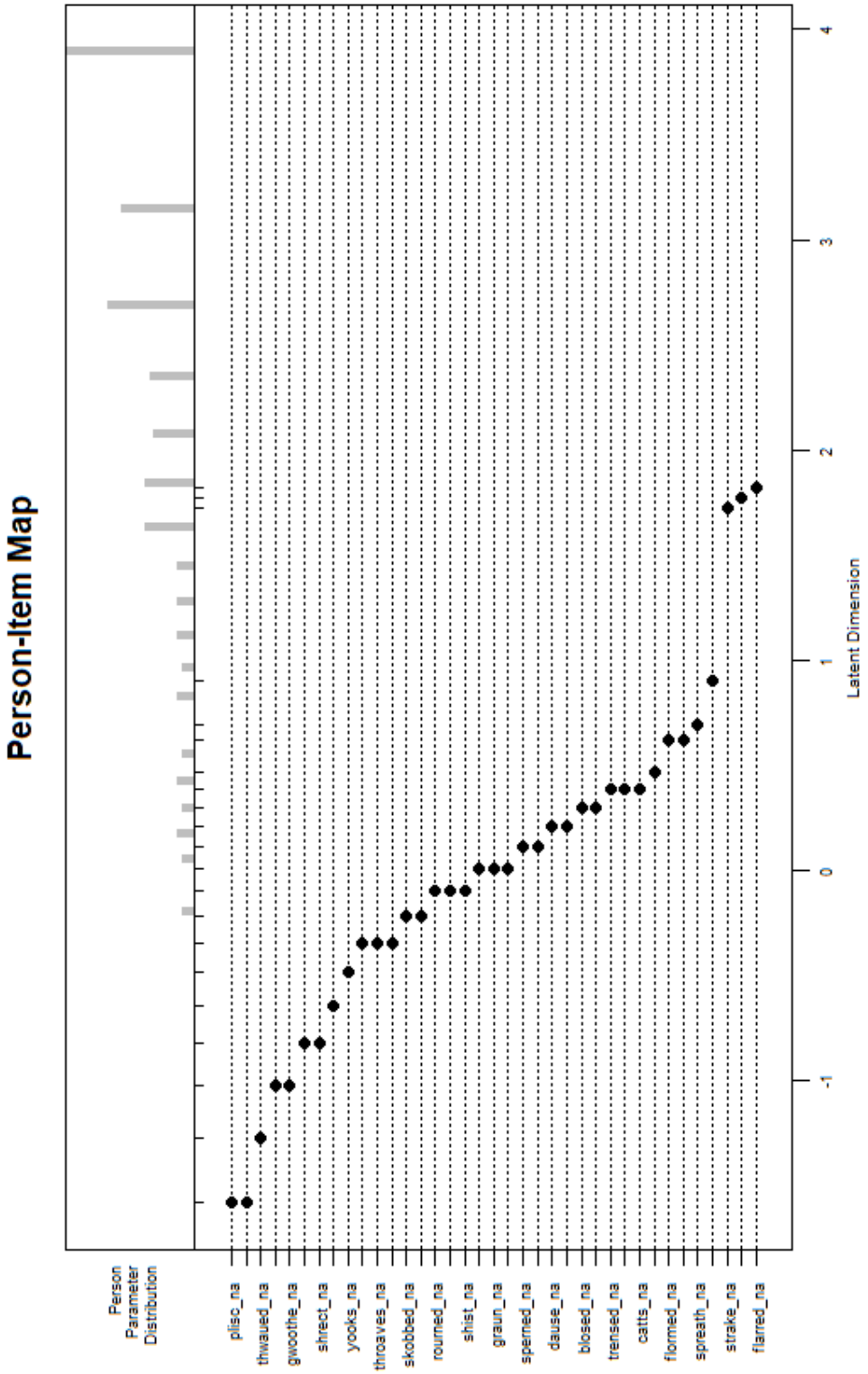


Figure 8
Person-Item Map for the Yes-No RT Test - Non-Word Data (37-item)



Masked Repetition Priming

In this task, 16 participants had a false alarm rate larger than 0.50, ranging from 0.53 – 0.91. These values suggested that they were either largely guessing on the non-words and/or did not have the proficiency levels that this study targeted. For this reason, I removed them from further analysis. The overall accuracy rates for all trials (words and non-words), word trials (critical and filler trials), and critical trials were 82% (*SD* = 38%), 85% (*SD* = 36%), and 90% (*SD* = 31%), respectively. The overall false alarm rate (incorrect responses to non-words) was 21% (*SD* = 41%).

In terms of item screening, I modeled the trimmed (300 ms < RT < 2500 ms), reciprocally transformed reaction times (-1/RT) using a mixed-effects model with maximal random effects (Barr et al., 2013) and with condition (related [0] *vs.* unrelated [1]) as the fixed effect. The iterative process of item inspection and model fitting suggested that the split-half reliability for the by-participant random slopes peaked when all items were retained. Since the goal here was to obtain a reliable measure of learner ability, I decided to use the full, 40-item data set to build the final model for this task. Descriptive statistics and the model summaries are presented in Table 21 and Table 22.

Table 21
Descriptive Statistics for the Masked Repetition Priming Task

	Mean RT in Millisecond (*SD*)		Split-half Reliability for By-Participant Random Slopes
	Related (e.g., patience – calm)	Unrelated (e.g., chestnut – calm)	
K = 40	654 (232)	686 (231)	.962
K = 39	642 (204)	681 (211)	.961

Table 22

Summary of Mixed Models - Masked Repetition Priming Task

	m1 (*K* = 40)	
Fixed effects	Estimate (*SE*)[a]	*t (p)*
Intercept	-1.66 (0.03)	-58.24 (<.001)
Condition	0.10 (0.01)	11.85 (<.001)
Random effects	Variance (*SD*)	Intercept-slope correlation
By-participant intercept	0.08 (0.29)	
By-item intercept	0.01 (0.08)	
By- participant slope for condition	0.001 (0.03)	-.97
By- item slope for condition	0.001 (0.03)	-.18
Residual	0.10 (0.2)	
Number of Observations	8271	

Semantic Priming

The by-participant accuracy analysis showed that 25 participants had a false alarm rate (incorrect responses to non-words) higher than 0.50, ranging from 0.51 – 0.98. I removed the data of these participants because of their high level of guessing on the non-words, casting doubt on their attention levels during the experiment and/or proficiency levels suitable for the task. The overall accuracy rates for all trials (words and non-words), word trials (critical and filler trials), critical trials (prime and target trials), and prime trials were 85% (*SD* = 36%), 87% (*SD* = 33%), 93% (*SD* = 26%), and 96% (*SD* = 19%), respectively. The false alarm rate (incorrect responses to non-words) was 21% (*SD* = 41%). In addition, it might be note-worthy that the accuracy for the prime trials (*M* = 89%, *SD* = 32%) was descriptively lower than the target trials (*M* = 96%, *SD* = 19%), which was expected because the prime words were the critical words in the study. Therefore, the priming effect can be attributed to knowledge of the prime words

(i.e., the critical words of the study) and did not depend on participants' knowledge of the target words in the experiment which elicited a ceiling accuracy.

In terms of item and reliability inspection, the full data set resulted in the highest level of reliability. I used this data set to build the mixed model from which I saved the by-participant random slopes as the measure of this task. Descriptive statistics and the model summary are presented in Table 23 and Table 24. Overall priming was not observed, meaning that, as a group, these participants did not appear to have robust semantic networks for this whole set of stimuli. Despite the high split-half reliability for the by-participant random slopes, the lack of priming casted doubt on what was being measured. For this reason, I discarded this measure in the following analysis.

Table 23

Means and Standard Deviations of Reaction Times for Critical Words Between Conditions (Semantic Priming)

	Mean RT in Millisecond (*SD*)		Split-half Reliability for By-Participant Random Slopes
	Related (e.g., patience – calm)	Unrelated (e.g., chestnut – calm)	
K = 40	615 (219)	622 (218)	.94
K = 39[a]	595 (172)	608 (185)	.85
K = 38[b]	605 (187)	612 (192)	.90

Note. [a] removing an item that elicited the most reverse-priming; [b] removing two items that elicited the most priming.

Table 24

Summary of Mixed Models - Semantic Priming Task

	m1 (*K* = 40)	
Fixed effects	Estimate (*SE*)[a]	*t (p)*
Intercept	-1.78 (0.03)	-55.33 (<.001)
Condition	0.02 (0.02)	1.03 (.31)
Random effects	Variance (*SD*)	Intercept-slope correlation
By-participant intercept	0.08 (0.27)	
By-item intercept	0.02 (0.12)	
By- participant slope for condition	0.001 (0.04)	.46
By- item slope for condition	0.02 (0.13)	-.64
Residual	0.10 (0.31)	
Number of Observations	7094	

Note. [a] all estimates were multiplied by 1000 for easier reading.

Summary of Results for Individual Tasks

In Table 25, I summarize the results for the individual tasks, including the number of items and participants included in the final data set, the descriptive statistics, the reliability estimates. I also present the correlation matrix for the task results in Table 26. These variables were standardized before submitted to the confirmatory factor analysis.

Table 25

Summary of Individual Task Results

Tasks	No. of Items	No. of Participants	Measure	Mean (*SD*)[a]	Reliability
Form-Meaning Receptive Test	35	144	Person-ability parameters from Rasch analysis	3.39 (1.27)	α = .89 ω = .90
Form-Meaning Productive Test	38	138	Person-ability parameters from Rasch analysis	1.48 (1.30)	α = .88 ω = .89
Yes-No RT Test (Accuracy)	word = 33; non-word = 37	138	Index of Signal Detection Theory	0.71 (0.22)	α = .85 (word) ω = .86 (word) α = .85 (non-word) ω = .85 (non-word)
Yes-No RT Test (Reaction Time)	40	137	Mean Reaction Time	772 (189)	Split-half = .87
Masked Repetition Priming	40	129	By-participant random slopes	1.5e-04 (0.03)	Split-half = .96
Semantic Priming	40	120	By-participant random slopes	7.4e-05 (0.02)	Split-half = .94

Note.[a] Empirical notation is used, where, for example, 1.2e-02 = 1.2 * 10^{-2} = 0.012

Table 26

Correlation Matrix for the Individual Task Results

	recep	prod	YesNoAcc	meanRT	RepPrim	SemPrim	prof
recep	1.00						
prod	0.57	1.00					
YesNoAcc	0.39	0.59	1.00				
meanRT	-0.36	-0.39	-0.36	1.00			
RepPrim	0.35	0.28	0.28	-0.55	1.00		
SemPrim	-0.17	-0.05	-0.11	0.39	-0.59	1.00	
prof	0.24	0.36	0.38	-0.24	0.25	-0.18	1.00

Notes. Recep = Form-Meaning Receptive Test; prod = Form-Meaning Productive Test; YesNoAcc = Yes-No RT Test (Accuracy); meanRT = Yes-No RT Test (Reaction Time); RepPrim = Masked Repetition Priming; SemPrim = Semantic Priming; prof = self-rated proficiency

CHAPTER 4: RESULTS (CFA AND SEM)

In this chapter, I report results of the confirmatory factor analysis and structural equation models which addressed my research questions. I will start with presenting the global model fit for all hypothesized models (i.e., the extent to which the model is an acceptable representation of the data), followed by an assessment of the local fit.

RQ1a – Explicit *vs.* Implicit

In Table 27, I present the level of fit of all the hypothesized models. Both one-factor solutions (with [CFA-M2] and without [CFA-M1] repetition priming) produced a good fit, suggesting evidence that lexical (implicit) knowledge measured by the repetition priming task can be placed on the same psychometric dimension as knowledge assessed by the other explicit tasks. Since these two models included different indicators, it was not appropriate to compare model fit between them. In terms of local fit, both models had standardized factor loadings ranging |0.50| to |0.91|, no standardized error variances larger than 1.96 (ranging from 0.17 to 0.79), and no modification indices larger than 3.84 (ranging from 0.03 to 2.70). I present the model summary of CFA-M2 in Table 28.

RQ1b – Knowledge *vs.* Strength

Similar to the one-factor, rivalry model (CFA-M2), the two-factor solution (CFA-M3) also produced a good fit (see fit indices in Table 27), suggesting evidence that lexical strength as assessed by timed tasks can be a psychometrically distinct dimension from lexical knowledge measured by untimed vocabulary tests. However, a X^2 difference test revealed no significant difference in the X^2 statistics (p = .243). For parsimony, the one-factor solution should be favored. Other fit indices also pointed to the same conclusion in that both the AIC and BIC had a

slightly lower value for the one-factor model than for the two-factor solution (AIC: 1732 *vs.* 1731; BIC: 1782 *vs.* 1779). In addition, the correlation between the two factors in CFA-M3 was at .915. This high correlation indicated that the knowledge measured by timed and untimed tasks was strongly related despite the potential, distinct dimensions. In terms of the local fit, no issues were found in the two-factor model. Standardized factor loadings ranged from |0.47| to |0.92|, no standardized error variances were larger than 1.96 (ranging from 0.16 to 0.78), and no modification indices were larger than 3.84 (ranging from 0.01 to 2.88). I present the model summary of CFA-M3 in Table 29.

Table 27

Summary of Confirmatory Factor Analysis and Structural Equation Models

Models	df	χ^2 (*p*)	RMSEA [95%CI]	CFI	SRMR	Fit	R^2 for Proficiency
Threshold for good fit		*p* < .05	< .05	> 0.95	< 0.08		
CFA-M1 (one-factor with four explicit indicators)	2	2.409 (.300)	0.038 [0.000, 0.174]	0.998	0.019	Good	
CFA-M2 (one-factors with all indicators)	4	4.915 (0.296)	0.040 [0.000, 0.137]	0.996	0.023	Good	
CFA-M3 (two-factors with knowledge and strength)	3	3.552 (0.314)	0.036 [0.000, 0.149]	0.998	0.022	Good	
SEM-M1a (CFA-M1 predicting proficiency)	5	4.609 (0.465)	0.000 [0.000, 0.111]	1.000	0.022	Good	0.261
SEM-M1b (CFA-M1 plus repprim predicting proficiency)	9	60.569 (< .001)	0.212 [0.164, 0.265]	0.763	0.166	Poor	
SEM-M2 (CFA-M2 predicting proficiency)	8	7.651 (0.468)	0.000 [0.000, 0.095]	1.000	0.025	Good	0.271
SEM-M3 (CFA-M3 predicting proficiency)	6	4.063 (0.668)	0.000 [0.000, 0.086]	1.000	0.019	Good	0.321
SEM-M4 (all five indictors predicting proficiency)				NA			0.190

Table 28

Model Summary of CFA-M2

Latent variables						
vocab =~		Estimate	Std.Err	z-value	P(>|z|)	Std.all
	recep	1.000				0.767
	prod	1.264	0.135	9.341	0.000	0.900
	YesNoAcc	0.963	0.116	8.309	0.000	0.730
	meanRT	-0.666	0.142	-4.703	0.000	-0.503
	repprim	0.611	0.117	5.243	0.000	0.456
Covariances:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
meanRT ~~						
	repprim	-0.375	0.088	-4.280	0.000	-0.471
Intercepts:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
	recep	0.000	0.083	0.000	1.000	0.000
	prod	-0.092	0.089	-1.029	0.303	-0.086
	YesNoAcc	-0.036	0.086	-0.416	0.677	-0.036
	meanRT	0.036	0.090	0.402	0.688	0.036
	repprim	-0.054	0.092	-0.592	0.554	-0.053
	vocab	0.000				0.000
Variances:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
	recep	0.409	0.060	6.806	0.000	0.411
	prod	0.218	0.072	3.038	0.002	0.190
	YesNoAcc	0.477	0.067	7.125	0.000	0.468
	meanRT	0.764	0.094	8.113	0.000	0.747
	repprim	0.829	0.136	6.080	0.000	0.792
	vocab	0.584	0.129	4.537	0.000	1.000

Notes. Recep = Form-Meaning Receptive Test; prod = Form-Meaning Productive Test; YesNoAcc = Yes-No RT Test (Accuracy); meanRT = Yes-No RT Test (Reaction Time); repprim = Masked Repetition Priming

Table 29

Model Summary of CFA-M3

Latent Variables:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
knowledge =~						
	recep	1.000				0.763
	prod	1.301	0.152	8.577	0.000	0.918
strength =~						
	YesNoAcc	1.000				0.781
	meanRT	-0.681	0.130	-5.224	0.000	-0.531
	repprim	0.606	0.126	4.824	0.000	0.467
Covariances:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
meanRT ~~						
	repprim	-0.354	0.09	-3.950	0.000	-0.458
knowledge ~~						
	strength	0.548	0.103	5.317	0.000	0.915
Intercepts:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
	recep	0.000	0.083	0.000	1.000	0.000
	prod	-0.095	0.090	-1.051	0.293	-0.088
	YesNoAcc	-0.035	0.086	-0.409	0.682	-0.035
	meanRT	0.036	0.090	0.399	0.690	0.035
	repprim	-0.051	0.092	-0.561	0.575	-0.050
	knowledge	0.000				0.000
	strength	0.000				0.000
Variances:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
	recep	0.415	0.062	6.704	0.000	0.418
	prod	0.183	0.079	2.307	0.021	0.157
	YesNoAcc	0.396	0.090	4.397	0.000	0.389
	meanRT	0.735	0.093	7.938	0.000	0.718
	repprim	0.816	0.140	5.813	0.000	0.782
	knowledge	0.578	0.131	4.407	0.000	1.000
	strength	0.622	0.135	4.609	0.000	1.000

Notes. Recep = Form-Meaning Receptive Test; prod = Form-Meaning Productive Test; YesNoAcc = Yes-No RT Test (Accuracy); meanRT = Yes-No RT Test (Reaction Time); repprim = Masked Repetition Priming

RQ2a – Predictive Validity of a Single Vocabulary Construct

When vocabulary was used as a single construct (with and without repetition priming) to predict self-rated proficiency, both SEMs (SEM-M1a & SEM-M2) produced a good fit (see Table 27). The regression path from vocabulary to proficiency was also significant in both models (b = 0.680, SE = 0.115, p < .001 in SEM-M1a; b = 0.680, SE = 0.114, p < .001 in SEM-M2). In terms of the explanatory power of these models in accounting for proficiency, the addition of the repetition priming to the vocabulary construct increased the R^2 value for self-rated proficiency from 0.261 in SEM-M1a to 0.271 in SEM-M2. This result suggested that the repetition priming task was not explaining much of the variance in proficiency. At the same time, this figure still compared favorably with that of the multiple regression model where the five measures served as the predictors (SEM-M4, R^2 = 0.190). Interestingly, when repetition priming was treated as a separate observed variable (predictor), the model (SEM-M1b) produced a poor fit, indicating model misspecification (i.e., relationships between variables in the data were not well reflected in the specification of the model). In other words, repetition priming was related to self-rated proficiency only when (or to a larger extent when) it was incorporated into a single vocabulary construct.

RQ2b – Predictive Validity of Lexical Knowledge and Strength

SEM-M3 where lexical knowledge and strength predicted proficiency produced a good fit (see Table 27). Both regression paths were not significant (knowledge -> proficiency: b = -0.167, SE = 0.637, p = .79; strength -> proficiency: b = 0.858, SE = 0.662, p = .195). This result indicated that neither lexical knowledge nor strength could account for unique variance in proficiency above and beyond each other. However, the R^2 value for self-rated proficiency

increased from 0.271 in SEM-M2 (with the one-factor solution as predictor) to 0.321 when knowledge and strength were treated as separate factors. In other words, the overall explanatory power was larger when timed and untimed measures were conceptualized as distinct constructs.

Summary of Findings

To summarize all findings, I found psychometric evidence that all five measures can belong to a single dimension, suggesting a uni-dimensional view of the vocabulary construct. At the same time, there are also signs indicating that lexical strength (as measured by timed tasks) can represent a distinct construct from lexical knowledge (as assessed untimed tests). When vocabulary is considered as separate constructs, its explanatory power for self-rated proficiency is strongest.

CHAPTER 5: DISCUSSION AND CONCLUSION

In this chapter, I discuss the findings in relation to the research questions as well as methodological issues pertaining the data analysis procedure. I will also suggest directions for further research before drawing a conclusion to close the present dissertation.

The Jury Is out but…

In this dissertation, I set out to validate six word measures by examining (1) the alignment between the psychological (e.g., explicit *vs.* implicit word knowledge) and psychometric dimensionality as well as (2) their predictive validity under different conceptualizations of the vocabulary construct. The present research represents the very first attempt in the field at extending vocabulary construct validation research to the domain of implicit *vs.* explicit and time sensitive *vs.* non-time sensitive word measures. In light of the present results, a straightforward answer is unlikely. On the contrary, the findings invite more research questions than they address.

In terms of the distinction between implicit and explicit word knowledge, the one-factor solution (CFA-M2) suggests evidence that the repetition priming task can be placed on the same psychometric dimension as the other explicit vocabulary measure, supporting a uni-dimensionality view of the vocabulary construct. However, in the absence of a two-factor model, the current findings remain silent on the extent to which there can be more than one dimension as far as implicit and explicit word knowledge is concerned. As for the distinction between lexical knowledge and strength (as measured by untimed and timed tasks, respectively), the picture is more complex. On the one hand, the two-factor model (CFA-M3) produces a good fit, suggesting that timed tasks can be considered as psychometrically distinct

from untimed tasks. On the other, the high correlation at .92 between the two constructs signals caution as it casts doubt on the discriminant validity of these factors. In addition, the rivalry, one-factor model (CFA-M2) fits equally well with the data, indicating that a one-factor representation of the data is also accurate and acceptable. For parsimony, the one-factor solution is preferred. Taken together, the present data set has provided evidence for a unitary view of vocabulary knowledge as judged by the factor structure. This view appears to hold true in both the implicit *vs*. explicit and the knowledge *vs.* strength distinctions.

A Broader, Unitary View of Vocabulary Knowledge

The good-fitting one-factor model (CFA-M2) points to the legitimacy to view vocabulary knowledge as a unitary construct if this finding is consistently replicated in the future. Although the present study is the first to include reaction time-based measures, this idea of a unitary view of vocabulary is not new. Focusing exclusively on accuracy measures of word component knowledge, González-Fernández and Schmitt (2020) also reported a one-factor model suggesting that one should view vocabulary as a single construct. In a similar vein, although Koizumi and In'nami (2020) reported a two-factor model for vocabulary size and depth, the two factors were highly correlated at .945, indicating that the two constructs (size and depth) are very highly related. It is reasonable then to suggest that this high correlation means a lack of discriminant validity between the two factors. Indeed, the authors' one-factor model also produced a good fit although the two-factor model fitted statistically better. In the present study, despite suggestions that timed measures would tap into distinct dimensions of lexical knowledge, I found no overwhelming evidence for this view in the present data set.

In qualifying their results, González-Fernández and Schmitt (2020) stress the interconnection between different word knowledge components. They also suggest that "no aspect is learned in a way that is detached from the other aspects" (González-Fernández & Schmitt, 2020, p. 498). Put differently, then, development in one aspect of word knowledge is likely to facilitate that of the other aspects. Similar claims may apply to the present context where explicit and implicit word knowledge as well as lexical knowledge and strength should be viewed as intimately connected. Although the present study remains silent on the potential developmental trajectory of individual aspects of the vocabulary construct, a unitary view can lead some to believe that they are acquired somewhat similarly. At least, acquisition of a certain aspect (e.g., explicit knowledge) should carry a facilitating role in the learning of other aspects (e.g., implicit knowledge). Indeed, Elgort (2011) show that intentional word learning can result in acquisition of word meaning that is measured by implicit tasks. A similar role of direct instruction in the development of implicit collocation knowledge is also reported by Toomer and Elgort (2019). Note that this view of similar developmental pathways for explicit and implicit word knowledge as well as lexical knowledge and strength differs from how implicit knowledge of grammar is believed to be acquired (i.e., it is learned dominantly through exposure, and explicit instruction only carries indirect roles). These differences highlight that, despite some parallelism, vocabulary and grammar research does not always mirror each other as far as explicit and implicit knowledge and learning are concerned.

Despite the unitary view as supported by the good fit for the one-factor solution, the association between the factor and the reaction time-based indicators is only of moderate strength. Standardized factor loadings were in the |.45| to |.50| range. This level of strength

signals that the factor has a broader coverage than what individual tasks measure, and hence a deviation from a good alignment between the factor and the reaction time measures. Essentially, as one loads more indicators to a factor, more sub-domain coverage is achieved. At the same time, the addition of indicators also changes the nature of the latent variable. This shift can cause a reduced strength of association between the factor and individual indicators because the factor's wider coverage often means that it is further away from what a specific task measures. On this account, the coverage of these five measures in the present study together represent multiple sub-domains of word knowledge that are all important to the conceptualization of vocabulary. Ignoring any of these aspects might result in an overly simplistic, narrow view of lexical knowledge. Therefore, researchers should answer the call for the use of measures that are of different natures (e.g., Elgort, 2018; Godfroid, 2020b; Vandenberghe et al., 2021). In doing so, more aspects of word knowledge can be accounted for in the vocabulary construct (see also discussion below on modeling vocabulary at a latent level). For example, as argued in Chapter 1, vocabulary tests administered without time pressure suffer from a lack of face validity (e.g., Godfroid, 2020b; Hui & Godfroid, 2020). To complement these untimed measures, researchers should start using more time-pressured tasks, especially when they are shown to provide additional information to traditional tests in the present study. This idea of obtaining further insights through implicit and timed measures leads the present discussion to the findings of the second research question regarding the predictive validity of the different conceptualizations of the vocabulary construct.

What Implicit and Timed Measures Offer

When paper- and accuracy-based vocabulary tests, which are relatively easy to administer, are already widely used in L2 research, a key question to ask is what additional information implicit and/or timed measures can offer to researchers. This issue bears practical implications on research practices, in addition to the theoretical discussion of the vocabulary construct (i.e., the psychological dimensionality). Given limited time and funds, should researchers only use explicit, untimed vocabulary tests? What is the value of the implicit and timed measures that is worth the resources? From my data, the addition of an implicit measure to the vocabulary construct (from CFA-M1 to CFA-M2) seems to increase the explanatory power of (self-reported) proficiency only to a very small extent. One straightforward interpretation may be that the repetition priming task is not really providing a lot more additional (statistical) information. At the same time, it is good to bear in mind that the repetition priming task targets the establishing of lexical representations. In other words, it is a task to examine the extent to which there is an entry in the mental lexicon for a given letter string. Even when detected, the lexical entries may or may not contain sufficiently enriched information (e.g., semantics) that can be put into authentic language use. On this account, the explanatory power of a vocabulary construct may also be related to the demand of the tasks in the battery as a whole. Indeed, language use (e.g., reading and listening) is often more highly correlated with a recall test than a recognition test (e.g., S. Zhang & Zhang, 2020), with the former placing more demand of knowledge on the learner. Therefore, it appears that the implicitness (or explicitness) of a task does not have a large impact upon the explanatory power of the vocabulary construct.

Another finding is that the level of predictive validity of the same set of measures depends also on the factor structure of the vocabulary construct. In particular, when vocabulary is conceptualized separately as lexical knowledge and strength (see CFA-M3), it has most explanatory power in accounting for the variance in proficiency. In this light, although the CFA results point to a one-factor model for parsimony, a two-factor solution based on time pressure can be more useful for researchers because it carries the kind of statistical information that can explain individual differences in proficiency among learners. In a way, the principle of parsimony should be an important factor to consider, but whether or not it should be the only criterion deserves some more deliberation. Potentially, researchers should take predictive validity into account when conceptualizing the vocabulary construct. In addition, the fact that different conceptualizations of the vocabulary construct have various levels of predictive power has implications on the construct validation research in grammar. Much, if not all, of the literature focuses on the factor structure of a battery of tests. A winner model is decided based on model fit and/or the principle of parsimony. Perhaps, examining the predictive validity of these conceptualizations could also lead to fruitful insights in grammar research as well.

Finally, it is worth noting that when these word measures are modeled at an observed level (SEM-M4), it has the least explanatory power. This result highlights the importance of modeling vocabulary at a latent level. First, CFA and SEM allow researchers to achieve more comprehensive sub-domain coverage. In the present study, CFA-M2, for example, has a single factor labelled as vocabulary. This factor has five indicators. In other words, this underlying construct of vocabulary covers all aspects that these five measures collectively tap into. With a more comprehensive coverage and the flexibility of specify the model according to different

theoretical conceptualizations, vocabulary knowledge as a latent variable can capture a learner's lexical proficiency to a fuller extent, as shown in the present findings. Therefore, this modeling of a latent variable goes one step beyond the mere use of multiple vocabulary measures in a study that researchers have called for (e.g., Read, 2020). Further, González-Fernández and Schmitt (2020) point out that using latent variables to examine vocabulary allows researchers to purify their vocabulary measure because relationships between variables (both latent and observed) are examined free of errors; therefore the representation (of vocabulary knowledge) is believed to be more accurate. Having a relatively pure measure of vocabulary is very important, especially when vocabulary is used to predict other outcomes of interest. When measurement errors are not properly modeling, they can bias the regression coefficient that most researchers are interested in. Last but not least, using SEM to study vocabulary allows researchers to have more flexibility to model the many different relationships related to vocabulary than regression analyses which can handle only one outcome. For example, one can investigate the extent to which a particular instruction method can lead to larger vocabulary growth, which in turn may enhance one's reading or listening performance. In this case, researchers can build a mediation model where treatment is the predictor, vocabulary is the mediator, and language performance is the outcome. It is through modeling these sophisticated relationships that researchers gain insights into how vocabulary is learned, and the ramifications of successful lexical development.

Understanding Priming Tasks as Individual Differences Measures

In addition to the discussion concerning the research questions, there are a couple of methodological notes. The first is related to the use of priming tasks in this line of research. As

discussed in Chapter 2, priming tasks have often been used in psycholinguistic experiments where researchers are interested in group-level effects. At the group level, if learners show priming, researchers conclude that the learners have established the relevant representations in memory. Therefore, oftentimes, results are interpreted in a binary fashion—either there is priming or there is no priming. In the present study, the priming tasks were used as individual differences measures. That is, the level of priming of an individual was used to index performance relative to the sample. This use of priming as an individual differences measure brought about two key issues: first, how the level and direction of priming are related to one's overall lexical knowledge; and second, how reliable priming tasks are when used in individual differences research.

When the level of priming is used as a continuous variable, it is necessary to understand what a high level of priming means. In other words, do researchers expect more skilled learners to demonstrate larger priming effects, for example? Or might it be the case that the opposite is predicted? These questions may not be very intuitive because priming tasks have mostly been used in contexts where researchers see priming as a binary. Indeed, my data show that, for the masked repetition task, higher levels of priming are associated with more skilled learners. This is manifested in two statistical estimates. First, in the analysis of the masked repetition priming task (Table 22), the mixed model estimated a negative correlation between the by-participant random intercepts and slopes at -.97. This means that when a participant has a low intercept value (representing overall faster responses in the related condition), they tend to have a high slope value (larger priming). Another estimate that signals more priming is better is in the CFA results. From Table 28, for example, it can be seen that the

standardized factor loading for the task is positive at 0.46, indicating that a learner scoring high on the factor (vocabulary knowledge) also scores high on the repetition priming task (larger priming).

However, the same observations do not apply to the semantic priming task. On the contrary, the opposite is true. For the semantic priming task, the more skilled a learner is, the less priming (or even reverse priming) is observed. Evidence can be found in correlation matrix (Table 26) where the correlation between the mean reaction time in the Yes-No RT test and semantic priming is positive at .39, representing that when a learner is slow (less skilled as manifested in a high RT), they demonstrate more priming. The semantic priming task also negatively correlates with the repetition priming task at -.59. In the mixed-effects model (Table 24), the correlation between the by-participant random intercepts and slopes is positive at .46, suggesting that when a learner is slow in the related condition (larger RT), they show more semantic priming. This is an unexpected finding that deserves further investigation. In addition, since overall priming is not observed, I was not confident to include this task in the further analysis because what was being measure is not entirely clear. In Elgort's (2011) study, the author notes that when the prime "is not fully acquired, the semantic priming effect may be inhibitory [reverse priming], whereas if the primes are acquired, the effect is facilitatory [priming]" (p. 394, my additions in brackets). Likewise, Bordag et al. (2015) took reverse priming (inhibition) as evidence of "memory traces of the new semantic representation" (p. 372). The lack of overall semantic priming for the present sample, therefore, may be a manifestation of opposite effects cancelling each other. That is, there was reverse priming for those who did not fully acquire all critical words, and there was some priming for those who did. Finally, more

proficient learners may also be more engaged in processing the semantics of the words, which causes them to respond slower.

Taken together, researchers should first clarify the characteristics of the semantic priming task by, for exampling, examining and confirming the fuller developmental trajectory of lexical knowledge as measured by a semantic priming task. Potentially, learners may first show no effects, followed by reverse priming (i.e., negative differences between reaction times in the related and the unrelated condition), and then priming (a positive difference). This U-shaped trajectory may somewhat resemble that of lexical processing stability where learners' processing initially becomes less stable when new representations are established, before it becomes more stable again (Hui, 2020). In addition, when researchers use a semantic priming task as an individual differences measure, they need to specify the direction of effects *a priori*; otherwise, claims of learning may be unfalsifiable as both priming and reverse priming can be taken as evidence for learning. Importantly, this non-linear trajectory of semantic priming can make the measure less reliable as noted by Elgort (2011).

Lastly, one initial concern for using priming tasks as an individual differences measure was their potential low reliability (Draheim et al., 2019). To mitigate this problem, I employed three strategies as discussed in Chapter 2: first, I used mixed-effects models to account for item-related variability (Rouder & Haaf, 2019). Second, I engaged in model criticism to treat outliers in fitting the mixed models (Baayen & Milin, 2010). Finally, I engaged in item inspection to attempt to remove random variability in the data. The first two strategies appear to be very useful in that the reliability levels of the two priming tasks are unexpectedly satisfactory (see Table 25). They are indeed on par with the accuracy-based measures. On the other hand, as

shown in Table 22 and Table 23, removing items does not always result in a (much) higher level of reliability. This success in achieving a relatively high reliability for priming measures has methodological implications for researchers using priming tasks as individual differences measures. Potentially, one can adopt these strategies and compare their impact on reliability with an aim to identify an optimal way for analyzing priming data. If this route is proven successful, one can assess if the same set of strategies can be applied to other processing time data, such as reading times obtained from eye tracking (Staub, 2021).

Alternative and Equivalent Models

Another methodological note concerns the fact that the good-fitting one- and two-factor models (CFA-M2 and CFA-M3) are statistically undistinguishable, suggesting that they are almost equally valid representations of the data. This scenario exemplifies a feature of confirmatory factor analysis (and indeed the structural equation modeling [SEM] framework in general) which is the existence of alternative and equivalent models. Brown (2015) defines equivalent solutions as "different model specifications produc[ing] identical goodness of fit (with the same number of df) and predicted covariance matrices (Σ) in any given data set" (p. 180). Since the fit for the models in the present case is not truly identical in statistical terms, I consider them to be alternative models. Indeed, both alternative and equivalent solutions are not uncommon. In an oft-cited paper, MacCallum et al. (1993) reviewed 20 articles using structural equation modeling that were published in a psychology journal between 1988 and 1991. MacCallum and colleagues (1993) found that all of them could have examined three or more equivalent models. The median number of potential equivalent models is three, with a range from three to 33,925. These numbers highlight the extent to which there exist potential,

alternative explanations to what is found by researchers. In terms of dealing with potential alternative and equivalent models, Brown (2015) suggests first ruling out theoretically implausible models, such as those that can be fitted but do not make practical sense (e.g., using data at Time 2 to predict performance at Time 1). In other cases, closely examining the contender models may have "considerable heuristic value" (Brown, 2015, p. 183). For example, researchers may not be aware of theoretical models that could also be plausible. Therefore, testing, comparing, and reporting alternative and equivalent models should be appreciated in any scientific endeavor.

Having said that, when there are competing theories, one approach researchers take may be to adjudicate on the case through rigorous, empirical work. I, myself, initially took this approach when designing this present study. But, I have also come to realize the limitations of having to decide on a final, best fitting, "winner" model. In particular, one could easily overlook what other potentially alternative and equivalent (or even less good fitting) models have to offer. On this account, it is unfortunate that researchers often do not acknowledge the existence of potentially alternative and equivalent models. In MacCallum et al. (1993), the authors reported that none of the articles they reviewed explicitly recognize the possibility of equivalent models, not to mention testing and comparing them. However, vocabulary researchers using confirmatory factor analysis (or SEM) have done a much better job in acknowledging alternative and equivalent models. For example, Koizumi and In'nami (2020) explicitly tested and compared different models as rivals of each other. González-Fernández and Schmitt (2020) also wrote that "[they] cannot claim that [their best fitting model] is the *only* valid statistical representation of vocabulary knowledge, but it is the model that best fit

[their] data, with its particular measures and participants" (p. 504, emphasis original). Taken together, I argue that researchers should take full advantage of alternative and equivalent models when they are found. In particular, these models allow researchers to examine the strengths and limitations of different conceptualizations, moving beyond a simplistic adjudication one might be the winner.

Limitations and Future Directions

First and foremost, this study needs to be replicated. In general terms, replication research can help researchers assess the extent to which findings reported are reliable. Types of replication research range from exact replications, where subsequent researchers repeat the study without any intended changes to the methodology, to partial replications, where one principled change is made, and to conceptual replications, where more changes are made (Marsden et al., 2018). While exact replications offer the most comparability between the initial and subsequent studies, partial replications can help researchers assess the generalizability of findings to other contexts, for instance. Therefore, the research field will benefit from researchers replicating this present study with, for instance, a different population (e.g., EFL learners) and/or with a different set of critical words and/or tasks.

Second, to the extent that the data allow, researchers should test more alternative and equivalent models to thoroughly understand the relationships between different aspects of vocabulary. For example, when a good two-factor model is available, one can assess if there is yet another second-order factor (e.g., vocabulary) governing the two first-order factors (explicit, and lexical strength and implicit word knowledge). Perhaps, a bi-factor model, where each indicator loads onto both a single factor (e.g., proficiency) and a construct factor (e.g.,

explicit word knowledge), can reveal the extent to which proficiency as a single factor drives performance in the indicators, in addition to explicit and implicit word knowledge. All these avenues present themselves as future directions for research.

Conclusion

This present study is one of the very first steps to shed light on the relationships between explicit and implicit as well as timed and untimed word measures. Using a battery of five vocabulary measures, I found evidence for a broad, unitary view of vocabulary knowledge. While these measures demonstrate psychometric unidimensionality, they represent a range of aspects of word knowledge that are collectively important to the construct of vocabulary. The initially hypothesized distinct dimension of lexical strength is not well supported by the present data, but this conceptualization offers the strongest predictive validity in explaining self-reported proficiency. Methodologically, I also demonstrated how priming data can be analyzed to achieve a high reliability level, a psychometric property that is particularly important to individual differences research.

Construct validity of measures carries important roles in (dis)confirming the theoretical conceptualizations of vocabulary as a construct. It is the prerequisite for language scientists who use quantitative methods to understand vocabulary learning and teaching. The dimensionality of word knowledge is as important as how researchers believe the knowledge is acquired and its implications on actual language performance. For example, could one ace all measures by exclusively studying word cards? Would naturalistic exposure (e.g., study abroad) promote implicit word knowledge more than explicit knowledge? What does it mean to have strong implicit word knowledge? Would strong implicit knowledge facilitate language

processing such as predictions that takes place during listening? To address these questions, researchers need to draw on measurement research such as this piece to make informed decisions on what vocabulary measures are appropriate for their research needs. In more general terms, vocabulary knowledge and learning should be modeled at the latent level to capture a more comprehensive, principled view of the construct.

APPENDICES

APPENDIX A

THE FORM-MEANING RECEPTIVE TEST

Instructions

See the closest meaning to the key word in the question. Here is an example.

SEE: They **saw** it.

a. cut

b. waited for

c. looked at

d. started

The answer is (c).

K2 Word Level

1. MAINTAIN: Can they **maintain** it?

a. keep it as it is

b. make it larger

c. get a better one than it

d. get it

2. STONE: He sat on a **stone**.

a. hard thing

b. kind of chair

c. soft thing on the floor

d. part of a tree

3. UPSET: I am **upset**.

a. tired

b. famous

c. rich

d. unhappy

4. DRAWER: The **drawer** was empty.

a. sliding box

b. place where cars are kept

c. cupboard to keep things cold

d. animal house

5. PATIENCE: He has no **patience**.

a. will not wait happily

b. has no free time

c. has no faith

d. does not know what is fair

6. CAP: They are talking about the cap.

a. cover for letters

b. kind of hat

c. place to live inside a tall building

d. food grown in garden

7. PUB: They went to the **pub**.

a. place where people drink and talk

b. place that looks after money

c. large building with many shops

d. building for swimming

8.CIRCLE: Make a **circle**.

a. rough picture

b. space with nothing in it

c. round shape

d. large hole

9. MICROPONE: Please use the **microphone**.

a. machine for making food hot

b. machine that makes sounds louder

c. machine that makes things look bigger

d. small telephone that can be carried around

10.PRO: He's a **pro**.

a. someone who is employed to find out important secrets

b. a stupid person

c. someone who writes for a newspaper

d. someone who is paid for playing sport etc.

K3 Word Level

1.SOLDIER: He is a **soldier**.

a. person in a business

b. student

c. person who uses metal

d. person in the army

2.RESTORE: It has been **restored**.

a. said again

b. given to a different person

c. given a lower price

d. made like new again

3. JUG: He was holding a **jug**.

a. a container for pouring liquids

b. an informal discussion

c. a soft cap

d. a weapon that explodes

4. SCRUB: He is **scrubbing** it.

a. cutting shallow lines into it

b. repairing it

c. rubbing it hard to clean it

d. drawing simple pictures of it

5.DINOSAUR: The children were pretending to be **dinosaurs**.

a. robbers who work at sea

b. very small creatures with human form but with wings

c. large creatures with wings that breathe fire

d. animals that lived a long time ago

Q16. STRAP: He broke the **strap**.

a. promise

b. top cover

c. shallow dish for food

d. strip of material for holding things together

Q17. PAVE: It was **paved**.

a. prevented from going through

b. divided

c. given gold edges

d. covered with a hard surface

Q18. DASH: They **dashed** over it.

a. moved quickly

b. moved slowly

c. fought

d. looked quickly

Q19. POVERTY: **Poverty** is a topic of discussion.

a. having little money

b. history

c. useful thing

d. action

Q20. LONESOME: He felt **lonesome**.

a. ungrateful

b. very tired

c. lonely

d. full of energy

K4 Word Level

Q31. COMPOUND: They made a new **compound**.

a. agreement

b. thing made of two or more parts

c. group of people forming a business

d. guess based on past experience

Q32. LATTER: I agree with the **latter**.

a. man from the church

b. reason given

c. last one

d. answer

33. CANDID: Please be **candid**.

a. be careful

b. show sympathy

c. show fairness to both sides

d. say what you really think

Q34. TUMMY: Look at my **tummy**.

a. cloth to cover the head

b. stomach

c. small furry animal

d. thumb

Q35. QUIZ: We made a **quiz**.

a. thing to hold arrows

b. serious mistake

c. set of questions

d. box for birds to make nests in

Q36. INPUT: We need more **input**.

a. information, power, etc. put into something

b. workers

c. artificial filling for a hole in wood

d. money

Q37. CRAB: Do you like **crabs**?

a. sea creatures that walk sideways

b. very thin small cakes

c. tight, hard collars

d. large black insects that sing at night

Q28.VOCABULARY: You will need more **vocabulary**.

a. words

b. skill

c. money

d. guns

Q29. REMEDY: We found a good **remedy**.

a. way to fix a problem

b. place to eat in public

c. way to prepare food

d. rule about numbers

Q30.ALLEGE: They **alleged** it.

a. claimed it without proof

b. stole the ideas for it from someone else

c. provided facts to prove it

d. argued against the facts that supported it

K5 Word Level

Q31. DEFICIT: The company had a large **deficit**.

a. spent a lot more money than it earned

b. went down a lot in value

c. had a plan for its spending that used a lot of money

d. had a lot of money in the bank

Q32. WEEP: He **wept**.

a. finished his course

b. cried

c. died

d. worried

Q33. NUN: We saw a **nun**.

a. long thin creature that lives in the earth

b. terrible accident

c. woman following a strict religious life

d. unexplained bright light in the sky

Q34. HAUNT: The house is **haunted**.

a. full of ornaments

b. rented

c. empty

d. full of ghosts

Q35. COMPOST: We need some **compost**.

a. strong support

b. help to feel better

c. hard stuff made of stones and sand stuck together

d. rotted plant material

Q36. CUBE: I need one more **cube**.

a. sharp thing used for joining things

b. solid square block

c. tall cup with no saucer

d. piece of stiff paper folded in half

Q37. MINIATURE: It is a **miniature**.

a. a very small thing of its kind

b. an instrument to look at small objects

c. a very small living creature

d. a small line to join letters in handwriting

Q38. PEEL: Shall I **peel** it?

a. let it sit in water for a long time

b. take the skin off it

c. make it white

d. cut it into thin pieces

Q39. FRACTURE: They found a **fracture**.

a. break

b. small piece

c. short coat

d. rare jewel

Q40. BACTERUM: They didn't find a single **bacterium**.

a. small living thing causing disease

b. plant with red or orange flowers

c. animal that carries water on its back

d. thing that has been stolen and sold to a shop

APPENDIX B

THE FORM-MEANING PRODUCTIVE TEST

Critical Words	Prompts
1. maintain	One needs to exercise regularly to mai_____ their fitness.
2. stone	When I was young, my father taught me how to skip a st____ across the pond.
3. upset	I didn't mean to up_____ him and make him cry – it was just a bit of fun.
4. drawer	If you pull out the dr_____, you can find the knives and forks.
5. patience	In the end, I lost my pat_____ and shouted at them.
6. cap	He likes wearing a baseball c___ because it can block the sunshine.
7. pub	They visit to go to the pu___ for a drink every Friday.
8. circle	If you want to teach a child to draw a sun, you can start by drawing a cir____.
9. microphone	People cannot hear well. Could you speak into the micr_____.
10. pro	She won nine games out of the ten that she played. She is surely a pr___.
11. soldier	At that point, the sol_____ in full uniform opened fire on the car.
12. restore	Although the house had a lot of damages, it has now been res__.
13. jug	She filled the ju____ up with milk.
14. scrub	He scr___ the dishes clean with a sponge.
15. dinosaur	Din_____ are a type of large animals that became extinct long ago.

16. strap	How can I adjust the str_____ on this helmet? It's too tight.
17. pave	Because of the holes on the road, they pa_____ it again.
18. dash	The dog ran off, and she da_____ after him.
19. poverty	Many people here live in po_____, making very little money for their living.
20. lonesome	Since there is no one around, he feels lones_____.
21. compound	A com_____ word is formed by putting two words together, like *classroom* is from the words *class* and *room*.
22. latter	Faced with two plans, she preferred the la____.
23. candid	If you say what you think, you are a can____ person.
24. tummy	Asian children are told not to have cold drinks or their tum_____ would hurt.
25. quiz	There was a pop qu____ in math at school today.
26. input	You don't need to thank me. My inp_____ into the project was just one idea.
27. crab	We went to a seafood restaurant, and we ordered cr___.
28. vocabulary	He needs to learn more words. A larger vo_____ can help with understanding the language.
29. remedy	Hot soup is the best rem____ for the common cold.
30. allege	Without any evidence, they all____ that man killed his neighbor.
31. deficit	The company is spending more than it makes. The owner is expecting a de_____ in the budget.

32. weep	After hearing the sad story, he wanted to we____.
33. nun	He was taught by Catholic n_____ at school. That's why he knows so much about the religion.
34. haunt	He was so scared walking out of the ha_____ house.
35. compost	Dead plants are used to create com_____ soil for gardening.
36. cube	Many Americans like to have ice cu____ in their water.
37. miniature	This one is too big. I like the mini_____ version.
38. peel	She doesn't like the skins. Would you p_____ these potatoes?
39. fracture	After the accident, the doctor saw a bone fra_____ in his x-ray film.
40. bacterum	After cleaning this with alcohol, I don't think we can find a single bact_____.

APPENDIX C

STIMULI FOR THE YES-NO RT TEST

ItemNo	Item	WordType	ItemNo	Item	WordType
01	maintain	word	41	yimb	nonword
02	stone	word	42	snarbs	nonword
03	upset	word	43	ghieved	nonword
04	drawer	word	44	knilms	nonword
05	patience	word	45	mulb	nonword
06	cap	word	46	shists	nonword
07	pub	word	47	thraifs	nonword
08	circle	word	48	psuse	nonword
09	microphone	word	49	dause	nonword
10	pro	word	50	zolved	nonword
11	soldier	word	51	yooks	nonword
12	restore	word	52	trensed	nonword
13	jug	word	53	twouche	nonword
14	scrub	word	54	flarred	nonword
15	dinosaur	word	55	rourned	nonword
16	strap	word	56	chift	nonword
17	pave	word	57	brisps	nonword
18	dash	word	58	graun	nonword
19	poverty	word	59	gwoothed	nonword

20	lonesome	word	60	spreathed	nonword
21	compound	word	61	flormed	nonword
22	latter	word	62	prarns	nonword
23	candid	word	63	shrut	nonword
24	tummy	word	64	skoign	nonword
25	quiz	word	65	broined	nonword
26	input	word	66	jeg	nonword
27	crab	word	67	skobbed	nonword
28	vocabulary	word	68	gnoathe	nonword
29	remedy	word	69	clersed	nonword
30	allege	word	70	spermed	nonword
31	deficit	word	71	blossed	nonword
32	weep	word	72	chead	nonword
33	nun	word	73	gwoaked	nonword
34	haunt	word	74	strake	nonword
35	compost	word	75	zatts	nonword
36	cube	word	76	melch	nonword
37	miniature	word	77	shrect	nonword
38	peel	word	78	throaves	nonword
39	fracture	word	79	thwagued	nonword
40	bacterium	word	80	plisc	nonword

APPENDIX D

STIMULI FOR THE MASKED REPETITION PRIMING TASK

ItemNo	Prime	Target	TargetType	TrialType	Condition
C01	maintain	MAINTAIN	word	critical	related
C01	announce	MAINTAIN	word	critical	unrelated
C02	stone	STONE	word	critical	related
C02	chest	STONE	word	critical	unrelated
C03	upset	UPSET	word	critical	related
C03	major	UPSET	word	critical	unrelated
C04	drawer	DRAWER	word	critical	related
C04	actors	DRAWER	word	critical	unrelated
C05	patience	PATIENCE	word	critical	related
C05	occasion	PATIENCE	word	critical	unrelated
C06	nil	CAP	word	critical	related
C06	lop	CAP	word	critical	unrelated
C07	pub	PUB	word	critical	related
C07	gap	PUB	word	critical	unrelated
C08	circle	CIRCLE	word	critical	related
C08	effect	CIRCLE	word	critical	unrelated
C09	microphone	MICROPHONE	word	critical	related
C09	suspension	MICROPHONE	word	critical	unrelated
C10	pro	PRO	word	critical	related

C10	shy	PRO	word	critical	unrelated
C11	soldier	SOLDIER	word	critical	related
C11	account	SOLDIER	word	critical	unrelated
C12	restore	RESTORE	word	critical	related
C12	cherish	RESTORE	word	critical	unrelated
C13	jug	JUG	word	critical	related
C13	icy	JUG	word	critical	unrelated
C14	scrub	SCRUB	word	critical	related
C14	draft	SCRUB	word	critical	unrelated
C15	dinosaur	DINOSAUR	word	critical	related
C15	triangle	DINOSAUR	word	critical	unrelated
C16	strap	STRAP	word	critical	related
C16	pause	STRAP	word	critical	unrelated
C17	pave	PAVE	word	critical	related
C17	loom	PAVE	word	critical	unrelated
C18	dash	DASH	word	critical	related
C18	memo	DASH	word	critical	unrelated
C19	poverty	POVERTY	word	critical	related
C19	dentist	POVERTY	word	critical	unrelated
C20	lonesome	LONESOME	word	critical	related
C20	concrete	LONESOME	word	critical	unrelated
C21	compound	COMPOUND	word	critical	related

C21	sympathy	COMPOUND	word	critical	unrelated
C22	latter	LATTER	word	critical	related
C22	unkind	LATTER	word	critical	unrelated
C23	candid	CANDID	word	critical	related
C23	unreal	CANDID	word	critical	unrelated
C24	tummy	TUMMY	word	critical	related
C24	poems	TUMMY	word	critical	unrelated
C25	quiz	QUIZ	word	critical	related
C25	lump	QUIZ	word	critical	unrelated
C26	input	INPUT	word	critical	related
C26	stool	INPUT	word	critical	unrelated
C27	crab	CRAB	word	critical	related
C27	bolt	CRAB	word	critical	unrelated
C28	vocabulary	VOCABULARY	word	critical	related
C28	psychiatry	VOCABULARY	word	critical	unrelated
C29	remedy	REMEDY	word	critical	related
C29	crater	REMEDY	word	critical	unrelated
C30	allege	ALLEGE	word	critical	related
C30	pitted	ALLEGE	word	critical	unrelated
C31	deficit	DEFICIT	word	critical	related
C31	surname	DEFICIT	word	critical	unrelated
C32	weep	WEEP	word	critical	related

C32	bark	WEEP	word	critical	unrelated
C33	nun	NUN	word	critical	related
C33	dip	NUN	word	critical	unrelated
C34	haunt	HAUNT	word	critical	related
C34	unite	HAUNT	word	critical	unrelated
C35	compost	COMPOST	word	critical	related
C35	leakage	COMPOST	word	critical	unrelated
C36	cube	CUBE	word	critical	related
C36	polo	CUBE	word	critical	unrelated
C37	miniature	MINIATURE	word	critical	related
C37	incorrect	MINIATURE	word	critical	unrelated
C38	peel	PEEL	word	critical	related
C38	echo	PEEL	word	critical	unrelated
C39	fracture	FRACTURE	word	critical	related
C39	database	FRACTURE	word	critical	unrelated
C40	bacterium	BACTERIUM	word	critical	related
C40	aggregate	BACTERIUM	word	critical	unrelated
F01	package	WOLF	word	filler	unrelated
F02	lion	LAKE	word	filler	unrelated
F03	stem	SIGN	word	filler	unrelated
F04	weapon	TRIPOD	word	filler	unrelated
F05	soil	BRAKE	word	filler	unrelated

F06	scab	WINDOW	word	filler	unrelated
F07	coast	SOCCER	word	filler	unrelated
F08	umpire	MERCURY	word	filler	unrelated
F09	string	FEET	word	filler	unrelated
F10	vest	COURT	word	filler	unrelated
F11	swimming	WINK	word	filler	unrelated
F12	clothes	WICK	word	filler	unrelated
F13	womb	PLATTER	word	filler	unrelated
F14	willow	RING	word	filler	unrelated
F15	rope	NOSE	word	filler	unrelated
F16	toilet	MOON	word	filler	unrelated
F17	wine	WIND	word	filler	unrelated
F18	sand	TONGUE	word	filler	unrelated
F19	rock	LOOT	word	filler	unrelated
F20	cockpit	MULE	word	filler	unrelated
F21	veil	RASH	word	filler	unrelated
F22	home	WOOD	word	filler	unrelated
F23	tunnel	ACCORDION	word	filler	unrelated
F24	cider	CHALK	word	filler	unrelated
F25	plum	HOOD	word	filler	unrelated
F26	panties	BURRO	word	filler	unrelated
F27	candy	NOTE	word	filler	unrelated

F28	coral	MENU	word	filler	unrelated
F29	sneeze	REED	word	filler	unrelated
F30	king	SHIP	word	filler	unrelated
F31	lung	CHEST	word	filler	unrelated
F32	roof	LUMP	word	filler	unrelated
F33	cabinet	CEDAR	word	filler	unrelated
F34	lime	MOSS	word	filler	unrelated
F35	brownie	MONARCH	word	filler	unrelated
F36	star	KNEE	word	filler	unrelated
F37	wren	SHOE	word	filler	unrelated
F38	hump	MOLE	word	filler	unrelated
F39	pork	MANSION	word	filler	unrelated
F40	nail	SULTAN	word	filler	unrelated
F41	soda	TIRE	word	filler	unrelated
F42	temple	RUBY	word	filler	unrelated
F43	sofa	MEAT	word	filler	unrelated
F44	tube	CHART	word	filler	unrelated
F45	mustard	STREET	word	filler	unrelated
F46	plug	CREAM	word	filler	unrelated
F47	ambulance	TOWN	word	filler	unrelated
F48	sketch	RAIL	word	filler	unrelated
F49	triangle	SPIDER	word	filler	unrelated

F50	supper	WOODLAND	word	filler	unrelated
F51	yellow	MILK	word	filler	unrelated
F52	channel	PEAR	word	filler	unrelated
F53	lamb	PINE	word	filler	unrelated
F54	pill	KISS	word	filler	unrelated
F55	pope	PUMP	word	filler	unrelated
F56	lice	TROMBONE	word	filler	unrelated
F57	twig	CHEEK	word	filler	unrelated
F58	lawn	COUCH	word	filler	unrelated
F59	jail	SISTER	word	filler	unrelated
F60	walnut	TOOL	word	filler	unrelated
F61	tortoise	VIOLET	word	filler	unrelated
F62	sycamore	LIMB	word	filler	unrelated
F63	aluminium	PASSAGE	word	filler	unrelated
F64	cigar	URCHIN	word	filler	unrelated
F65	bloom	TAIL	word	filler	unrelated
F66	oatmeal	WOOL	word	filler	unrelated
F67	ceiling	SIXPENCE	word	filler	unrelated
F68	scorpion	POLE	word	filler	unrelated
F69	lily	TURTLE	word	filler	unrelated
F70	rain	WALL	word	filler	unrelated
F71	sunset	TANK	word	filler	unrelated

F72	stable	CHAIN	word	filler	unrelated
F73	brother	SONG	word	filler	unrelated
F74	china	SINK	word	filler	unrelated
F75	pool	SURF	word	filler	unrelated
F76	hook	CRANE	word	filler	unrelated
F77	crowd	ROOT	word	filler	unrelated
F78	meal	LOCK	word	filler	unrelated
F79	ramp	JEEP	word	filler	unrelated
F80	hoof	CHOIR	word	filler	unrelated
N01	snarbs	SNARBS	non	non	related
N01	plisc	SNARBS	non	non	unrelated
N02	knilms	KNILMS	non	non	related
N02	lymphs	KNILMS	non	non	unrelated
N03	shists	SHISTS	non	non	related
N03	ficed	SHISTS	non	non	unrelated
N04	zolved	ZOLVED	non	non	related
N04	rhoiled	ZOLVED	non	non	unrelated
N05	ghieved	GHIEVED	non	non	related
N05	shrut	GHIEVED	non	non	unrelated
N06	brisps	BRISPS	non	non	related
N06	skoign	BRISPS	non	non	unrelated
N07	prarns	PRARNS	non	non	related

N07	frifth	PRARNS	non	non	unrelated
N08	dause	DAUSE	non	non	related
N08	blunch	DAUSE	non	non	unrelated
N09	graun	GRAUN	non	non	related
N09	clersed	GRAUN	non	non	unrelated
N10	shrut	SHRUT	non	non	related
N10	gnewth	SHRUT	non	non	unrelated
N11	skoign	SKOIGN	non	non	related
N11	zolved	SKOIGN	non	non	unrelated
N12	trensed	TRENSED	non	non	related
N12	yooks	TRENSED	non	non	unrelated
N13	twouche	TWOUCHE	non	non	related
N13	stroop	TWOUCHE	non	non	unrelated
N14	flarred	FLARRED	non	non	related
N14	zatched	FLARRED	non	non	unrelated
N15	strake	STRAKE	non	non	related
N15	shaifs	STRAKE	non	non	unrelated
N16	shrect	SHRECT	non	non	related
N16	stelks	SHRECT	non	non	unrelated
N17	broined	BROINED	non	non	related
N17	flarred	BROINED	non	non	unrelated
N18	zatts	ZATTS	non	non	related

N18	cronked	ZATTS	non	non	unrelated
N19	yimb	YIMB	non	non	related
N19	thwerge	YIMB	non	non	unrelated
N20	psuse	PSUSE	non	non	related
N20	clike	PSUSE	non	non	unrelated
N21	flunes	FLUNES	non	non	related
N21	thwagued	FLUNES	non	non	unrelated
N22	yooks	YOOKS	non	non	related
N22	ghieved	YOOKS	non	non	unrelated
N23	driped	DRIPED	non	non	related
N23	strake	DRIPED	non	non	unrelated
N24	spreathed	SPREATHED	non	non	related
N24	broined	SPREATHED	non	non	unrelated
N25	prives	PRIVES	non	non	related
N25	cless	PRIVES	non	non	unrelated
N26	gwilns	GWILNS	non	non	related
N26	chift	GWILNS	non	non	unrelated
N27	chift	CHIFT	non	non	related
N27	chead	CHIFT	non	non	unrelated
N28	chead	CHEAD	non	non	related
N28	prowse	CHEAD	non	non	unrelated
N29	melch	MELCH	non	non	related

N29	thraifs	MELCH	non	non	unrelated
N30	spreat	SPREAT	non	non	related
N30	phroaps	SPREAT	non	non	unrelated
N31	vaides	VAIDES	non	non	related
N31	spreathed	VAIDES	non	non	unrelated
N32	plisc	PLISC	non	non	related
N32	snarbs	PLISC	non	non	unrelated
N33	rhorts	RHORTS	non	non	related
N33	psuse	RHORTS	non	non	unrelated
N34	skobbed	SKOBBED	non	non	related
N34	knilms	SKOBBED	non	non	unrelated
N35	jumbed	JUMBED	non	non	related
N35	prives	JUMBED	non	non	unrelated
N36	snance	SNANCE	non	non	related
N36	farge	SNANCE	non	non	unrelated
N37	clersed	CLERSED	non	non	related
N37	spreat	CLERSED	non	non	unrelated
N38	stroop	STROOP	non	non	related
N38	truiff	STROOP	non	non	unrelated
N39	drongs	DRONGS	non	non	related
N39	pised	DRONGS	non	non	unrelated
N40	wrukes	WRUKES	non	non	related

N40	snooks	WRUKES	non	non	unrelated
N41	blunch	BLUNCH	non	non	related
N41	janns	BLUNCH	non	non	unrelated
N42	stelks	STELKS	non	non	related
N42	prathed	STELKS	non	non	unrelated
N43	gwoaked	GWOAKED	non	non	related
N43	trensed	GWOAKED	non	non	unrelated
N44	thraifs	THRAIFS	non	non	related
N44	shrect	THRAIFS	non	non	unrelated
N45	rourned	ROURNED	non	non	related
N45	gurve	ROURNED	non	non	unrelated
N46	dradge	DRADGE	non	non	related
N46	gwoothed	DRADGE	non	non	unrelated
N47	rhoiled	RHOILED	non	non	related
N47	vaides	RHOILED	non	non	unrelated
N48	psith	PSITH	non	non	related
N48	zatts	PSITH	non	non	unrelated
N49	truiff	TRUIFF	non	non	related
N49	psith	TRUIFF	non	non	unrelated
N50	gnewth	GNEWTH	non	non	related
N50	rhean	GNEWTH	non	non	unrelated
N51	frifth	FRIFTH	non	non	related

N51	flunes	FRIFTH	non	non	unrelated
N52	thwagued	THWAGUED	non	non	related
N52	driped	THWAGUED	non	non	unrelated
N53	rhean	RHEAN	non	non	related
N53	gwoaked	RHEAN	non	non	unrelated
N54	jeg	JEG	non	non	related
N54	gwilns	JEG	non	non	unrelated
N55	dweem	DWEEM	non	non	related
N55	rhorts	DWEEM	non	non	unrelated
N56	zatched	ZATCHED	non	non	related
N56	skurs	ZATCHED	non	non	unrelated
N57	zorns	ZORNS	non	non	related
N57	dradge	ZORNS	non	non	unrelated
N58	snooks	SNOOKS	non	non	related
N58	ghurrs	SNOOKS	non	non	unrelated
N59	thwerge	THWERGE	non	non	related
N59	snance	THWERGE	non	non	unrelated
N60	prathed	PRATHED	non	non	related
N60	dweem	PRATHED	non	non	unrelated
N61	janns	JANNS	non	non	related
N61	jumbed	JANNS	non	non	unrelated
N62	skurs	SKURS	non	non	related

N62	jeg	SKURS	non	non	unrelated
N63	phroaps	PHROAPS	non	non	related
N63	zorns	PHROAPS	non	non	unrelated
N64	gwoothed	GWOOTHED	non	non	related
N64	brisps	GWOOTHED	non	non	unrelated
N65	plurrs	PLURRS	non	non	related
N65	flormed	PLURRS	non	non	unrelated
N66	wharked	WHARKED	non	non	related
N66	melch	WHARKED	non	non	unrelated
N67	cless	CLESS	non	non	related
N67	shists	CLESS	non	non	unrelated
N68	frelt	FRELT	non	non	related
N68	wrukes	FRELT	non	non	unrelated
N69	prowse	PROWSE	non	non	related
N69	skobbed	PROWSE	non	non	unrelated
N70	gurve	GURVE	non	non	related
N70	wharked	GURVE	non	non	unrelated
N71	farge	FARGE	non	non	related
N71	prarns	FARGE	non	non	unrelated
N72	shaifs	SHAIFS	non	non	related
N72	zulbs	SHAIFS	non	non	unrelated
N73	cronked	CRONKED	non	non	related

N73	rourned	CRONKED	non	non	unrelated
N74	zulbs	ZULBS	non	non	related
N74	plurrs	ZULBS	non	non	unrelated
N75	ghurrs	GHURRS	non	non	related
N75	graun	GHURRS	non	non	unrelated
N76	clike	CLIKE	non	non	related
N76	yimb	CLIKE	non	non	unrelated
N77	lymphs	LYMPHS	non	non	related
N77	drongs	LYMPHS	non	non	unrelated
N78	flormed	FLORMED	non	non	related
N78	twouche	FLORMED	non	non	unrelated
N79	pised	PISED	non	non	related
N79	frelt	PISED	non	non	unrelated
N80	ficed	FICED	non	non	related
N80	dause	FICED	non	non	unrelated

APPENDIX E

WORD ASSOCIATION NORMS FOR CRITICAL RELATED TRIALS

ItemNo	Prime	Target	EAT	USF-FAN
C01	maintain	keep		
C02	stone	brick	0.02	
C03	upset	worried		
C04	drawer	desk	0.05	0.14
C05	patience	calm	0.05	0.17
C06	cap	head	0.19	0.11
C07	pub	bar	0.05	0.29
C08	circle	ball		
C09	microphone	voice		0.04
C10	pro	champion		
C11	soldier	military		0.07
C12	restore	build	0.01	0.04
C13	jug	pouring	0.01	
C14	scrub	cleaner		
C15	dinosaur	fossil		
C16	strap	bra	0.08	0.06
C17	pave	road		
C18	dash	race	0.01	
C19	poverty	hunger	0.05	

C20	lonesome	sad	0.03	
C21	compound	laboratory		
C22	latter	former	0.90	
C23	candid	honest		
C24	tummy	belly	0.02	
C25	quiz	exam		
C26	input	entry		
C27	crab	pinch		
C28	vocabulary	grammar		
C29	remedy	solution	0.02	0.09
C30	allege	accuse		
C31	deficit	budget		
C32	weep	sorrow		
C33	nun	church		
C34	haunt	ghost	0.50	0.32
C35	compost	soil		
C36	cube	square		
C37	miniature	tiny		
C38	peel	banana	0.01	0.13
C39	fracture	break	0.35	0.62
C40	bacterium	disease	0.08	0.13

Note. EAT_BW = Edinburgh Associative Thesaurus (Backward Association); EAT_FW = Edinburgh Associative Thesaurus (Forward Association); USF-FAN_BW = University of South Florida Free Association Norms (Backward Association); USF-FAN_FW = University of South Florida Free Association Norms (Forward Association)

APPENDIX F

STIMULI FOR THE MASKED SEMANTIC PRIMING TASK

ItemNo	Prime	Target	TargetType	TrialType	Condition	Cosine
C01	maintain	keep	word	critical	related	0.55
C01	announce	keep	word	critical	unrelated	0.85
C02	stone	brick	word	critical	related	0.61
C02	chest	brick	word	critical	unrelated	0.89
C03	upset	WORRIED	word	critical	related	0.34
C03	major	WORRIED	word	critical	unrelated	0.83
C04	drawer	DESK	word	critical	related	0.47
C04	actors	DESK	word	critical	unrelated	0.91
C05	patience	calm	word	critical	related	0.68
C05	chestnut	calm	word	critical	unrelated	0.95
C06	cap	head	word	critical	related	0.68
C06	lop	head	word	critical	unrelated	0.86
C07	pub	bar	word	critical	related	0.52
C07	gap	bar	word	critical	unrelated	0.98
C08	circle	ball	word	critical	related	0.74
C08	effect	ball	word	critical	unrelated	0.93
C09	microphone	voice	word	critical	related	0.56

C09	suspension	voice	word	critical	unrelated	0.95
C10	pro	champion	word	critical	related	0.67
C10	shy	champion	word	critical	unrelated	0.90
C11	soldier	military	word	critical	related	0.55
C11	account	military	word	critical	unrelated	0.98
C12	restore	build	word	critical	related	0.63
C12	cherish	build	word	critical	unrelated	0.85
C13	jug		word	critical	related	0.69
C13	icy		word	critical	unrelated	0.79
C14	scrub	CLEANER	word	critical	related	0.71
C14	draft	CLEANER	word	critical	unrelated	0.95
C15	dinosaur	fossil	word	critical	related	0.48
C15	triangle	fossil	word	critical	unrelated	1.08
C16	strap	bra	word	critical	related	0.64
C16	essay	bra	word	critical	unrelated	0.91
C17	pave	road	word	critical	related	0.76
C17	loom	road	word	critical	unrelated	0.92
C18	dash	race	word	critical	related	0.80
C18	memo	race	word	critical	unrelated	0.94
C19	poverty	hunger	word	critical	related	0.54
C19	dentist	hunger	word	critical	unrelated	0.95
C20	lonesome	SAD	word	critical	related	0.61

C20	concrete	SAD	word	critical	unrelated	0.87
C21	compound	LABORATORY	word	critical	related	0.71
C21	sympathy	LABORATORY	word	critical	unrelated	0.99
C22	latter	FORMER	word	critical	related	0.67
C22	unkind	FORMER	word	critical	unrelated	0.92
C23	candid	HONEST	word	critical	related	0.59
C23	unreal	HONEST	word	critical	unrelated	0.82
C24	tummy	belly	word	critical	related	0.54
C24	poems	belly	word	critical	unrelated	0.90
C25	quiz	exam	word	critical	related	0.70
C25	lump	exam	word	critical	unrelated	0.86
C26	input	entry	word	critical	related	0.85
C26	stool	entry	word	critical	unrelated	1.02
C27	crab	pinch	word	critical	related	0.88
C27	haze	pinch	word	critical	unrelated	0.88
C28	vocabulary	grammar	word	critical	related	0.52
C28	psychiatry	grammar	word	critical	unrelated	0.94
C29	remedy	solution	word	critical	related	0.59
C29	crater	solution	word	critical	unrelated	0.94
C30	allege	ACCUSE	word	critical	related	0.56
C30	pitted	ACCUSE	word	critical	unrelated	0.85
C31	deficit	BUDGET	word	critical	related	0.59

C31	surname	BUDGET	word	critical	unrelated	1.06
C32	weep	SORROW	word	critical	related	0.47
C32	bark	SORROW	word	critical	unrelated	0.83
C33	nun	CHURCH	word	critical	related	0.65
C33	dip	CHURCH	word	critical	unrelated	0.95
C34	haunt	GHOST	word	critical	related	0.55
C34	unite	GHOST	word	critical	unrelated	0.90
C35	compost	SOIL	word	critical	related	0.55
C35	leakage	SOIL	word	critical	unrelated	0.90
C36	cube	SQUARE	word	critical	related	0.70
C36	polo	SQUARE	word	critical	unrelated	1.01
C37	miniature	TINY	word	critical	related	0.59
C37	incorrect	TINY	word	critical	unrelated	0.97
C38	peel	banana	word	critical	related	0.56
C38	echo	banana	word	critical	unrelated	0.93
C39	fracture	BREAK	word	critical	related	0.70
C39	database	BREAK	word	critical	unrelated	1.02
C40	bacterium	disease	word	critical	related	0.62
C40	aggregate	disease	word	critical	unrelated	1.06
F01	home	phial	word	filler	unrelated	0.86
F02	garbage	bow	word	filler	unrelated	0.86

F03	suds	cell	word	filler	unrelated	0.98
F04	bronze	trail	word	filler	unrelated	0.98
F05	ramp	slush	word	filler	unrelated	0.86
F06	latch	camp	word	filler	unrelated	1.02
F07	flood	servant	word	filler	unrelated	0.95
F08	steam	carpet	word	filler	unrelated	0.87
F09	chicken	gingham	word	filler	unrelated	0.79
F10	truck	emerald	word	filler	unrelated	1.03
F11	farmyard	casement	word	filler	unrelated	0.91
F12	stub	ticket	word	filler	unrelated	0.75
F13	branch	squire	word	filler	unrelated	1.01
F14	oboe	bank	word	filler	unrelated	0.99
F15	building	cologne	word	filler	unrelated	0.86
F16	abscess	bill	word	filler	unrelated	0.87
F17	aunt	lettuce	word	filler	unrelated	0.87
F18	pollen	mixer	word	filler	unrelated	0.96
F19	bourbon	rope	word	filler	unrelated	0.87
F20	pedal	hoe	word	filler	unrelated	0.88
F21	fellow	nickel	word	filler	unrelated	0.80

F22	boy	soap	word	filler	unrelated	0.80
F23	tweezer	horn	word	filler	unrelated	0.93
F24	lane	crow	word	filler	unrelated	0.82
F25	head	velvet	word	filler	unrelated	0.83
F26	rake	velvet	word	filler	unrelated	0.85
F27	necklace	prairie	word	filler	unrelated	0.89
F28	dance	sleet	word	filler	unrelated	0.90
F29	men	forearm	word	filler	unrelated	0.88
F30	pocket	shutter	word	filler	unrelated	0.89
F31	coke	missile	word	filler	unrelated	0.88
F32	blade	worker	word	filler	unrelated	0.90
F33	material	kernel	word	filler	unrelated	0.91
F34	dweller	bread	word	filler	unrelated	0.79
F35	tennis	hood	word	filler	unrelated	0.89
F36	fan	stable	word	filler	unrelated	1.00
F37	ship	cloak	word	filler	unrelated	0.84
F38	soup	shell	word	filler	unrelated	0.79
F39	bagpipe	naval	word	filler	unrelated	0.93
F40	spire	student	word	filler	unrelated	1.02

F41	seaman	paint	word	filler	unrelated	0.97
F42	thong	pole	word	filler	unrelated	0.80
F43	ammonia	malaria	word	filler	unrelated	0.95
F44	trapeze	brownie	word	filler	unrelated	0.93
F45	harvest	lumber	word	filler	unrelated	0.82
F46	china	yew	word	filler	unrelated	0.86
F47	shield	kerchief	word	filler	unrelated	0.82
F48	straw	music	word	filler	unrelated	0.95
F49	vault	cinnamon	word	filler	unrelated	0.99
F50	blouse	mantle	word	filler	unrelated	0.76
F51	mole	ether	word	filler	unrelated	0.89
F52	drill	novel	word	filler	unrelated	0.99
F53	male	child	word	filler	unrelated	0.76
F54	quarter	doctor	word	filler	unrelated	0.92
F55	bible	cottage	word	filler	unrelated	0.94
F56	epistle	pancreas	word	filler	unrelated	0.89
F57	boulder	hog	word	filler	unrelated	0.93
F58	shed	bristle	word	filler	unrelated	0.97
F59	nun	larch	word	filler	unrelated	0.95

F60	corpse	sofa	word	filler	unrelated	0.77
F61	slime	lynx	word	filler	unrelated	0.98
F62	plane	land	word	filler	unrelated	0.64
F63	machine	referee	word	filler	unrelated	0.97
F64	wife	guest	word	filler	unrelated	0.72
F65	puddle	walrus	word	filler	unrelated	0.95
F66	mineral	ankle	word	filler	unrelated	0.93
F67	emperor	crumb	word	filler	unrelated	0.90
F68	elephant	cafe	word	filler	unrelated	0.92
F69	rash	brim	word	filler	unrelated	1.02
F70	oil	crowd	word	filler	unrelated	0.96
F71	ramrod	corridor	word	filler	unrelated	0.81
F72	drain	rung	word	filler	unrelated	0.93
F73	dump	pea	word	filler	unrelated	0.92
F74	banana	beak	word	filler	unrelated	0.82
F75	toy	umpire	word	filler	unrelated	0.90
F76	cage	tailor	word	filler	unrelated	0.90
F77	statue	dough	word	filler	unrelated	0.88
F78	diving	body	word	filler	unrelated	0.87

F79	janitor	halter	word	filler	unrelated	0.99
F80	weapon	magician	word	filler	unrelated	0.77
N01	shryst	dweased	non	non	na	na
N02	juits	whilns	non	non	na	na
N03	spurved	theethe	non	non	na	na
N04	smighs	phleuds	non	non	na	na
N05	thralked	glurch	non	non	na	na
N06	pseague	skiln	non	non	na	na
N07	cloob	skaved	non	non	na	na
N08	guins	pheeped	non	non	na	na
N09	thriest	phrirts	non	non	na	na
N10	dweet	speemed	non	non	na	na
N11	tord	jalt	non	non	na	na
N12	swant	pralled	non	non	na	na
N13	snam	gheche	non	non	na	na
N14	sproque	blibbed	non	non	na	na
N15	plulf	blild	non	non	na	na
N16	swants	gooms	non	non	na	na
N17	psurnt	ufts	non	non	na	na

N18	strurse	farged	non	non	na	na
N19	clitch	wrursts	non	non	na	na
N20	smarps	kolts	non	non	na	na
N21	plowl	frouts	non	non	na	na
N22	gwaths	skeld	non	non	na	na
N23	dwerd	chowth	non	non	na	na
N24	snawse	cloot	non	non	na	na
N25	swegg	thafes	non	non	na	na
N26	cumbed	blulls	non	non	na	na
N27	swerch	drithed	non	non	na	na
N28	phromped	tudged	non	non	na	na
N29	sneefs	keph	non	non	na	na
N30	sliche	shrur	non	non	na	na
N31	gidge	sporde	non	non	na	na
N32	spreese	stumes	non	non	na	na
N33	yamps	rast	non	non	na	na
N34	prage	ghaitch	non	non	na	na
N35	thourged	dwagged	non	non	na	na
N36	thrilth	zimes	non	non	na	na

N37	yighed	yoal	non	non	na	na
N38	phrombed	wared	non	non	na	na
N39	zirms	stroards	non	non	na	na
N40	thogs	myed	non	non	na	na
N41	trume	wheamed	non	non	na	na
N42	shilch	chur	non	non	na	na
N43	spriel	tudd	non	non	na	na
N44	glalc	smaunch	non	non	na	na
N45	knund	ormed	non	non	na	na
N46	frant	flusk	non	non	na	na
N47	tafts	fomb	non	non	na	na
N48	plaired	rooths	non	non	na	na
N49	scrauve	snerts	non	non	na	na
N50	momes	spocs	non	non	na	na
N51	rhurb	deich	non	non	na	na
N52	ghised	hersed	non	non	na	na
N53	knawn	wrass	non	non	na	na
N54	croints	kear	non	non	na	na
N55	blaled	loamed	non	non	na	na

N56	taive	gwoured	non	non	na	na
N57	shrarps	dwades	non	non	na	na
N58	pryp	pryths	non	non	na	na
N59	theezed	ghooze	non	non	na	na
N60	clyes	whoy	non	non	na	na
N61	crisk	yeathed	non	non	na	na
N62	cloothe	splurf	non	non	na	na
N63	thweined	chowls	non	non	na	na
N64	trowse	thwelt	non	non	na	na
N65	skuned	ghinked	non	non	na	na
N66	skask	troots	non	non	na	na
N67	gnanch	smuids	non	non	na	na
N68	dwek	clake	non	non	na	na
N69	ghoathed	twirped	non	non	na	na
N70	knirr	gneke	non	non	na	na
N71	zarc	splee	non	non	na	na
N72	walds	stupe	non	non	na	na
N73	shroved	skulged	non	non	na	na
N74	spligned	sperk	non	non	na	na

N75	gief	truv	non	non	na	na
N76	wrarned	stryth	non	non	na	na
N77	shoign	luilds	non	non	na	na
N78	toathed	knuch	non	non	na	na
N79	kands	cagues	non	non	na	na
N80	maunt	flevved	non	non	na	na

REFERENCES

REFERENCES

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

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

Andringa, S., & Rebuschat, P. (2015). New directions in the study of implicit and explicit learning: An introduction. *Studies in Second Language Acquisition*, *37*(2), 185–196. https://doi.org/10.1017/S027226311500008X

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01501-5

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 026553222092748. https://doi.org/10.1177/0265532220927487

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

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.21500/20112084.807

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235–274. https://doi.org/10.1177/026553220101800301

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

Bodner, G. E., & Masson, M. E. J. (2001). Prime validity affects masked repetition priming: Evidence for an episodic resource account of priming. *Journal of Memory and Language*, *45*(4), 616–647. https://doi.org/10.1006/jmla.2001.2791

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. https://doi.org/10.1017/S1366728914000078

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.

Brysbaert, M. (2020). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition*, Advance online publication. https://doi.org/10.1017/S1366728920000437

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. https://doi.org/10.3758/BRM.41.4.977

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272. https://doi.org/10.1017/S0267190599190135

Chapelle, C. A. (2021). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11–20). Routledge.

Chen, Y. (2021). Comparing incidental vocabulary learning from reading-only and reading-while-listening. *System*, *97*, 102442. https://doi.org/10.1016/j.system.2020.102442

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3–25. https://doi.org/10.1177/0265532216676851

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editor's introduction. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modeling and assessing vocabulary knowledge* (pp. 1–32). Cambridge University Press.

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & Ml. H. Long (Eds.), *Handbook of second language acquisition* (pp. 313–348). Blackwell Publishing Ltd.

Denovan, A., & Dagnall, N. (2019). Development and evaluation of the Chronic Time Pressure Inventory. *Frontiers in Psychology*, *10*, 2717. https://doi.org/10.3389/fpsyg.2019.02717

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508–535. https://doi.org/10.1037/bul0000192

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Elgort, I. (2017). Incorrect inferences and contextual word learning in English as a second language. *Journal of the European Second Language Association*, *1*(1), 1. https://doi.org/10.22599/jesla.3

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. https://doi.org/10.1017/S0272263117000109

Elgort, I., & Piasecki, A. E. (2014). The effect of a bilingual learning mode on the establishment of lexical semantic representations in the L2. *Bilingualism: Language and Cognition*, *17*(3), 572–588. https://doi.org/10.1017/S1366728913000588

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables: L2 vocabulary learning from reading. *Language Learning*, *64*(2), 365–414. https://doi.org/10.1111/lang.12052

Ellis, N. C. (1994). Consciousness in second language learning: Psychological perspectives on the role of conscious processes in vocabulary acquisition. *AILA Review*, *11*, 37–56.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, *27*(02), 305–352. https://doi.org/10.1017/S027226310505014X

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning*, *54*(2), 227–275. https://doi.org/10.1111/j.1467-9922.2004.00255.x

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*(02), 141–172. https://doi.org/10.1017/S0272263105050096

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*(01). https://doi.org/10.1017/S0272263107070052

Evett, L. J., & Humphreys, G. W. (1981). The use of abstract graphemic information in lexical access. *The Quarterly Journal of Experimental Psychology*, *33*(4), 325–350. https://doi.org/10.1080/14640748108400797

Forster, K. I. (1998). The pros and cons of masked priming. *Journal of Psycholinguistic Research*, *27*(2), 203–233. https://doi.org/10.1023/A:1023202116609

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680–698. https://doi.org/10.1037/0278-7393.10.4.680

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Godfroid, A. (2020a). *Eye Tracking in Second Language Acquisition and Bilingualism: A Research Synthesis and Methodological Guide* (1st ed.). Routledge. https://doi.org/10.4324/9781315775616

Godfroid, A. (2020b). Sensitive measures of vocabulary knowledge and processing: Expanding Nation's framework. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). https://doi.org/10.4324/9780429291586-28

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*, *21*(3), 563–584. https://doi.org/10.1017/S1366728917000219

Gollan, T. H., Forster, K. I., & Frost, R. (1997). Translation priming with different scripts: Masked priming with cognates and noncognates in hebrew-english bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1122–1139. https://doi.org/10.1037/0278-7393.23.5.1122

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Grainger, J., Diependaele, K., Spinelli, E., Ferrand, L., & Farioli, F. (2003). Masked repetition and phonological priming within and across modalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1256–1269. https://doi.org/10.1037/0278-7393.29.6.1256

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. https://doi.org/10.1080/17470218.2015.1038280

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, *35*(3), 423–449. https://doi.org/10.1017/S0272263113000041

Harrington, M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook*, *6*(1), 147–168. https://doi.org/10.1075/eurosla.6.10har

Harrington, M. (2018). *Lexical facility: Size, recognition speed and consistency as dimensions of second language vocabulary knowledge*. Palgrave Macmillan.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1–11. https://doi.org/10.1177/026553229200900102

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. https://doi.org/10.1017/S0272263199002089

Hox, J., J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel Analysis Techniques and Applications*. Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hui, B. (2020). Processing variability in intentional and incidental word learning: An extension of Solovyeva and Dekeyser (2018). *Studies in Second Language Acquisition*, *42*(2), 327–357. https://doi.org/10.1017/S0272263119000603

Hui, B., & Godfroid, A. (2020). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, Advance online publication. https://doi.org/10.1017/S0142716420000193

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227–245. https://doi.org/10.1191/0265532202lt229oa

Hulstijn, J. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in Second Language Acquisition*, *27*(02). https://doi.org/10.1017/S0272263105050084

Hulstijn, J. (2007). Psycholinguistic perspectives on language and its acquisition. In J. Cummins & C. Davison (Eds.), *The international handbook of English language teaching* (pp. 783–796). Springer.

Isemonger, I. M. (2007). Operational definitions of explicit and implicit knowledge: Response to R. Ellis (2005) and some recommendations for future research in this area. *Studies in Second Language Acquisition*, *29*(01). https://doi.org/10.1017/S0272263107070040

Issa, B. I., Faretta–Stutenberg, M., & Bowden, H. W. (2020). Grammatical and lexical development during short-term study abroad: Exploring l2 contact and initial proficiency. *The Modern Language Journal*, *104*(4), 860–879. https://doi.org/10.1111/modl.12677

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. (1999). Testing processing explanations for the asymmetry in masked cross-language priming. *Bilingualism: Language and Cognition*, *2*(1), 59–75. https://doi.org/10.1017/S1366728999000152

Jiang, N. (2013). *Conducting reaction time research in second language studies*. Routledge. https://doi.org/10.4324/9780203146255

Jiang, N. (2015). Six decades of research on lexical representation and processing in bilinguals. In J. W. Schwieter (Ed.), *The Cambridge Handbook of Bilingual Processing* (pp. 29–84). Cambridge University Press. https://doi.org/10.1017/CBO9781107447257.002

Kim, H. S., Lee, J. H., & Lee, H. (2020). The relative effects of L1 and L2 glosses on L2 learning: A meta-analysis. *Language Teaching Research*, 136216882098139. https://doi.org/10.1177/1362168820981394

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 368–390. https://doi.org/10.1207/s15328007sem1203_2

Kim, K. M., & Godfroid, A. (2019). Should we listen or read? Modality effects in implicit and explicit knowledge. *Modern Language Journal*, *103*(3), 648–664. https://doi.org/10.1111/modl.12583

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Baileu, & N. Hamilton-Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh University Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Koizumi, R., & In'nami, Y. (2020). Structural equation modeling of vocabulary size and depth using conventional and Bayesian methods. *Frontiers in Psychology*, *11*, 618. https://doi.org/10.3389/fpsyg.2020.00618

Kroll, J., & Tokowicz, N. (2005). Models of bilingual representation and processing: Looking back and to the future. In J. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 531–553). Oxford University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126–132). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-12396-4_12

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Leow, R. P. (2001). Attention, awareness, and foreign language behavior. *Language Learning*, *51*, 113–155. https://doi.org/10.1111/j.1467-1770.2001.tb00016.x

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Loewen, S., & Gönülal, T. (2015). Exploratory factor analysis and primcipal componetns analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 182–212). Routledge.

Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *The Modern Language Journal*, *105*(1), 187–193. https://doi.org/10.1111/modl.12700

Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2020). performance: Assessment of Regression Models Performance. *CRAN*. https://doi.org/10.5281/zenodo.3952174

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*(1), 185–199. https://doi.org/10.1037/0033-2909.114.1.185

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. https://doi.org/10.1017/S0272263119000615

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9). http://www.jstatsoft.org/v20/i09

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field: replication in second language research. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. https://doi.org/10.1177/0265532211430367

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press. http://site.ebrary.com/id/10163350

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. http://dx.doi.org.proxy1.cl.msu.edu/10.1037/met0000144

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 558–572. https://doi.org/10.1037/0278-7393.24.3.558

Meara, P. (1992). Network structures and vocabulary acquisition in a foreign language. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 62–70). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-12396-4_6

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. John Benjamins Pub. Co.

Meara, P. (2010). *EFL vocabulary tests* (2nd ed.). Centre for Applied Language Studies, University College Swansea.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–151.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Milton, J., & Fitzpatrick, T. (2014). *Dimensions of vocabulary knowledge*. Palgrave.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. https://doi.org/10.1191/0265532206lt321oa

Müller, M. (2020). Item fit statistics for Rasch analysis: Can we trust them? *Journal of Statistical Distributions and Applications*, *7*(1), 5. https://doi.org/10.1186/s40488-020-00108-7

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Nakata, T., & Elgort, I. (2020). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, 026765832092776. https://doi.org/10.1177/0267658320927764

Nation, I. S. P. (2012). *The BNC/COCA word family lists. Unpublished paper. Available at: Http://www.victoria.ac.nz/lals/about/staff/paul-nation*.

Nation, I. S. P. (2013a). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P. (2013b). *Learning vocabulary in another language* (Second Edition). Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary* (1st ed). Heinle, Cengage Learning.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. http://www.usf.edu/FreeAssociation/

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*(1), 09. https://doi.org/10.18806/tesl.v11i1.623

Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second language vocabulary learning. In A. Pavlenko (Ed.), *The bilingual mental lexicon* (pp. 125–160). Multilingual Matters. https://doi.org/10.21832/9781847691262-008

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489–509. https://doi.org/10.1177/0265532212438053

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly*, *53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*(4), 583–621. https://doi.org/10.1177/0267658319828413

Qian, D. D., & Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 66–80). Routledge. https://doi.org/10.4324/9780429291586-5

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

Ramezanali, N., Uchihara, T., & Faez, F. (2021). Efficacy of multimodal glossing on second language vocabulary learning: A meta-analysis. *TESOL Quarterly*, *n/a*(n/a), Advance online publication. https://doi.org/10.1002/tesq.579

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362. https://doi.org/10.1080/02724980244000099

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, *79*(1), 200–210. https://doi.org/10.1177/0013164417725127

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. https://doi.org/10.1177/026553229301000308

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209–227). John Benjamins.

Read, J. (2020). Key issues in measuring vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., pp. 545–560). Routledge. https://doi.org/10.4324/9780429291586-34

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Révész, A., & Brunfaut, T. (2021). Validating assessments for research purposes. In P. Winke & T. Brunfaut (Eds.), *The routledge handbook of second language acquisition and language testing* (pp. 21–32). Routledge. https://doi.org/10.4324/9781351034784

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, *26*, 452–467. https://doi.org/10.3758/s13423-018-1558-y

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. http://www.rstudio.com/

Ruiz, S., Chen, X., Rebuschat, P., & Meurers, D. (2019). Measuring individual differences in cognitive abilities in the lab and on the web. *PLOS ONE*, *14*(12), e0226217. https://doi.org/10.1371/journal.pone.0226217

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1–17. https://doi.org/10.1037/0096-1523.3.1.1

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*(3), 369–385. https://doi.org/10.1017/S0142716400010845

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). IAP Information Age Publishing.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*(1), 66–78.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, *63*(1), 121–159. https://doi.org/10.1111/j.1467-9922.2012.00730.x

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*(3), 723–751. https://doi.org/10.1111/lang.12129

Staub, A. (2021). How reliable are individual differences in eye movements in reading? *Journal of Memory and Language*, *116*, 104190. https://doi.org/10.1016/j.jml.2020.104190

Stenneken, P., Conrad, M., & Jacobs, A. M. (2007). Processing of syllables in production and recognition tasks. *Journal of Psycholinguistic Research*, *36*(1), 65–78. https://doi.org/10.1007/s10936-006-9033-8

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, *29*(4), 471–488. https://doi.org/10.1177/0265532211433033

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*(5), 1229–1261. https://doi.org/10.1017/S014271641700011X

Tanabe, M. (2016). Measuring second language vocabulary knowledge using a temporal method. *Reading in a Foreign Language*, *28*(1), 118–142.

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, *52*(6), 2372–2382. https://doi.org/10.3758/s13428-020-01389-1

Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, *69*(2), 405–439. https://doi.org/10.1111/lang.12335

Trofimovich, P., & McDonough, K. (Eds.). (2011). *Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics*. John Benjamins Pub. Co.

Ullman, M. T. (2001). The Declarative/Procedural Model of Lexicon and Grammar. *Journal of Psycholinguistic Research*, *30*(1), 37–69. https://doi.org/10.1023/A:1005204207369

Vafaee, P., & Kachinske, I. (2019). The inadequate use of confirmatory factor analysis in second language acquisition validation studies. *Studies in Applied Linguistics and TESOL*, Vol. 19 No. 2 (2019). https://doi.org/10.7916/SALT.V19I2.4184

Vafaee, 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. https://doi.org/10.1017/S0272263119000676

Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*(1), 59–95. https://doi.org/10.1017/S0272263115000455

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390–416. https://doi.org/10.1111/lang.12105

VanPatten, B., & Jegerski, J. (Eds.). (2014). *Research methods in second language psycholinguistics*. Routledge.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(01), 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2012). Depth of vocabulary knowledge. In C. A. Chapelle (Ed.), *The encyclopedia of Applied Linguistics* (pp. 1656–1663). Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal1325

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13–40. https://doi.org/10.3138/cmlr.53.1.13

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ*, *3*, e1058. https://doi.org/10.7717/peerj.1058

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370–371.

Yanagisawa, A., & Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 371–386). Routledge. https://doi.org/10.4324/9780429291586-24

Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading?: A meta-regression analysis. *Studies in Second Language Acquisition*, *42*(2), 411–438. https://doi.org/10.1017/S0272263119000688

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, 136216882091399. https://doi.org/10.1177/1362168820913998

Zhang, X., Liu, J., & Ai, H. (2020). Pseudowords and guessing in the Yes/No format vocabulary test. *Language Testing*, *37*(1), 6–30. https://doi.org/10.1177/0265532219862265