ABSTRACT

Title of Dissertation:	DOES MODALITY MATTER? AURAL AND WRITTEN VOCABULARY IN SECOND LANGUAGE LISTENING AND READING COMPREHENSION
	Takehiro Iizuka, Doctor of Philosophy, 2024
Dissertation directed by:	Assistant Professor Bronson Hui, Second Language Acquisition

This study examined the significance of the mode of delivery—aural versus written—in second language (L2) vocabulary knowledge and L2 comprehension skills. One of the unique aspects of listening comprehension that sets it apart from reading comprehension is the mode of delivery—language input is delivered not visually but aurally. Somewhat surprisingly, however, this difference has not always been considered, and in fact L2 listening studies are more often accompanied by written tests (of, e.g., vocabulary knowledge) than by aural tests. Few studies have systematically examined the impact of modality on comprehension skills and linguistic variables such as vocabulary either, despite the long-standing view of language skills being multimodal. In this study, therefore, I first examined the degree to which aural and written vocabulary is separate constructs. Then I examined how each of those constructs explains listening and reading comprehension skills differently. By using latent variable modeling, I also addressed limitations in previous studies, including undue influence from measurement error and unique characteristics of particular tests.

One hundred eighty-five adult Japanese learners of English took four aural and four written English vocabulary tests, with parallel test formats across the modalities to allow for comparison. The effect of words was averaged out by counterbalancing eight property-matched sets of words. The participants also took listening and reading comprehension tests. The dimensionality of vocabulary knowledge was examined by comparing one-factor and multi-factor models. The unique contribution of aural and written vocabulary knowledge to listening and reading comprehension was evaluated by latent variable path analysis. The difference in the sizes of aural and written vocabulary knowledge was examined by latent means modeling.

The results of the study were nuanced. Modality effects were observed in the sense that (1) a two-factor model of vocabulary knowledge with aural and written vocabulary had a significantly better fit to the data than a one-factor model, (2) aural vocabulary knowledge uniquely explained some variance in listening comprehension skills, and (3) the participants' aural vocabulary size was significantly smaller than their written vocabulary size. However, the effects of modality were limited in the sense that (1) the aural and written vocabulary knowledge factors were very highly correlated and (2) the common part of the two factors—general vocabulary knowledge—explained much more variance in each of listening and reading comprehension skills than modality-specific knowledge. These results suggest that, although aural versus written test modality effects do seem to exist in L2 vocabulary knowledge and comprehension skills, its practical impact is small compared with that of general vocabulary knowledge at least in the context where words are presented in isolation as in the present study.

DOES MODALITY MATTER? AURAL AND WRITTEN VOCABULARY IN SECOND LANGUAGE LISTENING AND READING COMPREHENSION

by

Takehiro Iizuka

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

Advisory Committee:
Assistant Professor Bronson Hui, Chair
Professor Emeritus Robert DeKeyser
Professor Kira Gor
Dr. Martyn Clark
Professor Gregory Hancock, Dean's Representative

Dedication

To my mother, Mayumi

Acknowledgements

First and foremost, I would like to express my immense gratitude to Dr. Bronson Hui, the chair of my dissertation committee, for his flexibility, patience, and guidance throughout this dissertation project. At the early stage of this project when I was feeling daunted by the largeness of the task, he told me to "eat an elephant (dissertation) one bite at a time." I have been keeping that in mind all along. Thank you, Bronson, for your continued support while I was eating this elephant! I am also extremely grateful to Dr. Robert DeKeyser. I learned a lot from him through three courses and two qualifying papers, which I am certain have lasting impact on me as an SLA researcher beyond graduation. I will also never forget his caring personality. I am also very thankful to Dr. Kira Gor and Dr. Martyn Clark. Their constructive feedback based on their expertise made me realize what I overlooked, and improved the quality of this work. Additionally, I am so grateful to Dr. Gregory Hancock for offering invaluable statistics courses and office hours. I watched his lecture videos over and over again and I heavily drew on his techniques in this project. I also greatly admire his excellent teaching ability and sincere attitude towards each individual. Many thanks also to Jason Struck, Sanshiroh Ogawa, Meghan Hersh, Shayna Bell, Justin Thetford, and Mitchell Smith for their help with research material development and/or test scoring, and to Yuichi Suzuki, Etsuo Taguchi, Ellen Scattergood, John Rippey, and many others for their help with participant recruitment.

This dissertation project was financially supported by our SLA program as well as the following external grants: the Duolingo English Test Doctoral Dissertation Award, the TOEFL Grant for Doctoral Research in Language Assessment, and the *Language Learning* Dissertation Grant. I am very thankful for these funding resources as they allowed me to purchase a license

for statistical software, recruit raters to calculate interrater reliability, and gain a larger sample of participants than otherwise possible.

Lastly, I would like to express my appreciation to the people who might not have been involved in this dissertation project directly but supported me in important ways. Some of these people are no longer in this physical world but with me in my heart. Thank you, Dr. Kimi Nakatsukasa, for inviting me into the exciting field of SLA research. Thank you, Dr. Mike Long, for offering intellectually stimulating lectures and generous feedback on my work. I am also thankful to my friends and colleagues in and out of the SLA program and my students in Japanese classes, who made my life in the States way more than just doing research. Finally, I would like to express my deepest appreciation to my family, who always offer me unconditional love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1: Background

1.1 Introduction

One of the unique aspects of listening comprehension that sets it apart from reading comprehension is the mode of delivery—language input is delivered not visually but aurally. Somewhat surprisingly, however, this difference has not always been considered, and in fact second language (L2) listening studies are more often accompanied by written tests (of, e.g., vocabulary knowledge) than by aural tests (Zhang & Zhang, 2022). Although the number of studies that include aural tests has been increasing in recent years (e.g., Hui & Godfroid, 2021; Vafaee & Suzuki, 2020), the number is still small, and few studies have specifically examined the effect of modality across listening and reading (but see, e.g., Cheng & Matthews, 2018; Milton et al., 2010). The present study examined the effects of modality on the assessment of L2 vocabulary knowledge, one of the most important components of linguistic knowledge for L2 listening comprehension (Mecartty, 2000; Vafaee & Suzuki, 2020). Furthermore, the study examined the extent to which the modality effects in vocabulary knowledge relate to L2 listening and reading comprehension skills. With structural equation modeling (SEM), this study attempted to tease apart modality effects from general vocabulary knowledge.

Note that the present study dealt with *adult* L2 learning and therefore in this dissertation, unless otherwise noted, it can be assumed that the participants in the studies cited were adults.

1.2 Multidimensionality of Vocabulary Knowledge

Vocabulary knowledge has long been considered multidimensional by researchers (e.g., Cheng & Matthews, 2018; Cronbach, 1942; González-Fernández, 2022; Qian, 2002). Currently,

the most influential framework appears to be Nation's (2013) taxonomy (see Table 1). In this framework, vocabulary knowledge is considered to have three major aspects: form, meaning, and use. It is further assumed that each of these major aspects can be divided into three components, namely, spoken, written, and word parts for form; form and meaning, concepts and referents, and associations for meaning; and grammatical functions, collocations, and constraints on use for use. Each of these components then can be seen in terms of receptive and productive knowledge as described in Table 1. The present study zoomed in on the dimensionality of spoken versus written form focusing on form-meaning connection and including both receptive and productive knowledge.

In the following sections, I will first present an overview of the mechanism of listening comprehension and then discuss the potential significance of the dimension of spoken versus written vocabulary knowledge in comprehension skills.

Table 1

Nation's (2013, p. 49) Taxonomy of Vocabulary Knowledge

		Receptive knowledge	Productive knowledge
Form	Spoken	What does the word sound like?	How is the word pronounced?
	Written	What does the word look like?	How is the word written and spelled?
	Word parts	What parts are recognisable in this word?	What word parts are needed to express the meaning?
Meaning	Form and meaning	What meaning does this word form signal?	What word form can be used to express this meaning?
	Concept and referents	What is included in the concept?	What items can the concept refer to?
	Associations	What other words does this make us think of?	What other words could we use instead of this one?
Use	Grammatical functions	In what patterns does the word occur?	In what patterns must we use this word?
	Collocations	What words or types of words occur with this one?	What words or types of words must we use with this one?
	Constraints on use	Where, when, and how often would we expect to meet this word?	Where, when, and how often can we use this word?

1.3 What Makes L2 Listening Comprehension Possible? The Big Picture

Second language listening comprehension is a complex process that is influenced by many factors, including but clearly not limited to, linguistic knowledge, processing ability and speed, cognitive skills, and background knowledge (Vandergrift, 2007). Before narrowing down the scope of research, it is, therefore, perhaps useful to see the big picture. A study by Andringa et al. (2012) is a good place to start the discussion as it is one of the most comprehensive studies with many measures, and they analyzed the data holistically as opposed to separately for each element (e.g., vocabulary and grammar). In particular, the researchers created four factors to predict listening comprehension skills: linguistic knowledge, processing speed, working memory, and reasoning ability. The knowledge factor was indexed by accuracy of vocabulary, grammar, and segmentation tasks, the processing speed factor was indexed by reaction time in several tasks such as word-monitoring and self-paced listening, and the working memory and reasoning ability factors were indexed by span tasks and IQ test, respectively. Both native (n = 121) and nonnative (n = 113) speakers of Dutch completed these tasks as well as a listening comprehension test. With SEM, the results showed that, while linguistic knowledge and processing speed were the significant predictors for the native group, linguistic knowledge and reasoning ability were the significant predictors for the nonnative group. For the nonnative group the standardized regression coefficients from linguistic knowledge and reasoning ability to listening comprehension were .95 and .26, respectively, suggesting that L2 listening comprehension skills are mostly a function of linguistic knowledge and they are also somewhat influenced by reasoning ability.

Similar findings were obtained in a study by Wallace (2022). The researcher examined the contributions of linguistic knowledge (indexed by vocabulary), background knowledge, metacognitive awareness, working memory, and attentional control, to L2 listening comprehension for Japanese learners of English ($N = 226$) using SEM. The results showed that linguistic knowledge, background knowledge, and attentional control had direct effects on listening comprehension, and metacognitive awareness had an indirect effect on comprehension via background knowledge. The strongest predictor was linguistic knowledge with a standardized effect of .67, clearly larger than those of the other predictors (.27 for background knowledge, .18 for attentional control, and .15 for metacognitive awareness).

Having broad scope, these studies suggest that linguistic knowledge is the primary determinant of successful L2 listening comprehension, although other factors such as cognitive skills and background knowledge also affect comprehension to some degree.

1.4 A Theoretical Model of L2 Listening Comprehension

If linguistic knowledge is important, how exactly does it contribute to listening comprehension? Field (2013) provided a theoretical model on that. As described in Figure 1, Field's (2013) model of listening comprehension includes three levels of processing to reach the propositional level of understanding (i.e., literal comprehension). At the first level, sounds are transformed into representations that conform to the phonological system of the language being spoken (e.g., phonemes, syllables). This stage is supported by the listener's phonological knowledge. At the second level, spoken word forms are searched in the listener's mental lexicon to find the best matches for the acoustic information. This stage is supported by the listener's lexical knowledge. At the third level, a syntactic structure is imposed on a group of spoken word

forms to interpret meaning beyond single words. This stage is supported by the listener's syntactic knowledge. After literal understanding is achieved in this way, interpretation is adjusted by other factors such as contextual information and pragmatic knowledge. On a side note, although Field (2013) described listening process sequentially, he also acknowledged that it does not necessarily mean that one level of processing waits upon another (as indicated by the reverse arrows in Figure 1).

Field (2013) provided a useful theoretical model that described the importance of knowledge of phonology, vocabulary, and grammar in listening comprehension, but he did not mention the relative significance of these linguistic knowledge components and for that matter it would be useful to look at empirical studies of L2 listening comprehension with multiple variables of linguistic knowledge.

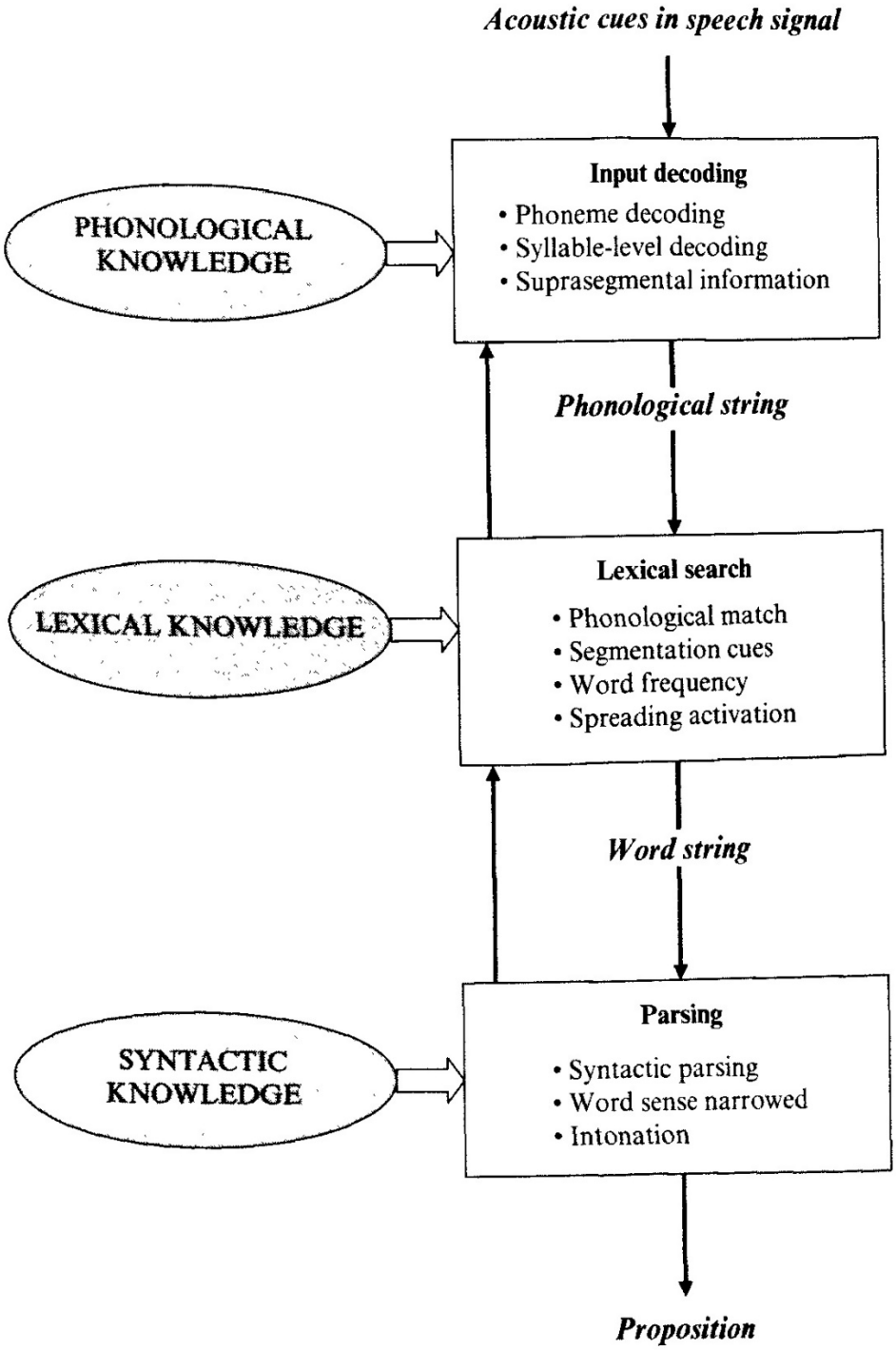


Figure 1. Field's (2013, p. 97) model of listening comprehension.

1.5 The Role of Linguistic Knowledge Components in L2 Listening Comprehension

Linguistic knowledge components have been examined in L2 listening comprehension. Mecartty (2000), for example, had 77 foreign language learners of Spanish take tests of vocabulary, grammar, and listening comprehension. Vocabulary knowledge was assessed by multiple-choice meaning recognition tests, and grammar knowledge was assessed by a grammaticality judgment test and a multiple-choice sentence completion test. The results showed that both vocabulary and grammar knowledge was moderately positively correlated with listening comprehension skills (r = .38 and .26, respectively). However, a multiple regression analysis showed that vocabulary knowledge was the only significant predictor when both vocabulary and grammar were entered into a model to predict listening comprehension, suggesting the importance of vocabulary knowledge in L2 listening comprehension.

Hui and Godfroid (2021) also examined knowledge of vocabulary and grammar in L2 listening comprehension for Chinese learners of English (N = 44). Vocabulary knowledge was assessed by a yes-no test in which the learners indicated if they knew the meaning of the words they heard. Grammar knowledge was assessed by a sentence construction test in which the learners selected a grammatical continuation of an aural sentence fragment. Aside from a general listening comprehension test, the researchers also included an aural sentence verification test where the learners judged the plausibility of sentences to assess the learners' listening skills at the propositional level (i.e., literal understanding of a sentence). The data was analyzed not only in terms of accuracy but also processing speed (reaction time) and automaticity (coefficient of variation). Multiple regression analyses showed that general listening comprehension skills were predicted by the accuracy of the literal understanding of sentences, and in turn the accuracy of the literal understanding was predicted by the accuracy and processing speed of vocabulary

knowledge. A subsequent mediation analysis showed that both accuracy and processing speed of vocabulary knowledge seemed to have impact on general listening comprehension skills, which was partially mediated by the accuracy of literal understanding. Grammar knowledge was not a significant predictor at any stage. These results suggest that vocabulary knowledge plays a fundamental role in L2 listening comprehension.

Vafaee and Suzuki (2020), on the other hand, examined the relative significance of vocabulary and grammar knowledge in L2 listening comprehension for Iranian learners of English ($N = 263$). Vocabulary knowledge was assessed by multiple-choice vocabulary breadth and depth tests. Grammar knowledge was assessed by a grammaticality judgment test and a sentence comprehension test with target syntactic structures that were expected to be particularly important for listening comprehension. Metacognitive awareness, working memory, and anxiety were also included as control variables. With SEM, the researchers showed that both vocabulary and grammar knowledge significantly contributed to listening comprehension. The standardized regression coefficients, however, were rather different between these two predictors: .55 for vocabulary knowledge versus .28 for grammar knowledge, underscoring the importance of vocabulary knowledge in L2 listening comprehension.

In addition, Wong et al. (2017) examined various types of phonological knowledge on top of vocabulary knowledge in L2 listening comprehension. The phonological variables included the ability to perceive phonologically reduced forms (a dictation task), minimal pair discrimination skills, the ability to immediately recognize a word by hearing only the initial part of the word (a gating task), and phonemic awareness. Vocabulary knowledge was assessed by a multiple-choice test where the participants selected synonyms of target words. A test of phonological short-term memory was also included. Sixty Chinese learners of English completed

these tasks and a listening comprehension test. Multiple regression analysis identified two significant predictors of listening comprehension skills: the reduced form dictation and the vocabulary test. The reduced form dictation score, one of the phonological variables, appeared to be the strongest predictor with a standardized regression coefficient of .40 (vs. .26 for the vocabulary test), but the dictation task not only required phonological knowledge but also vocabulary knowledge and so it is difficult to discuss here the relative significance of phonological and lexical knowledge. In fact, another regression analysis showed that the scores on the reduced form dictation were explained by the scores on the vocabulary test and the gating task.

Overall, the studies reviewed in this section support Field's (2013) model of listening comprehension in the sense that knowledge of phonology, vocabulary, and grammar are all important to some extent. At the same time, these empirical studies seem to suggest that vocabulary knowledge is the key variable that makes L2 listening comprehension possible.

1.6 The Role of Linguistic Knowledge Components in L2 Reading Comprehension

At this point, it may be useful to see the case of L2 *reading* comprehension also. For reading, there is a meta-analysis conducted by Jeon and Yamashita (2014). They synthesized 67 independent samples and calculated the average correlations between L2 reading comprehension skills and various variables. After adjusting for measurement error, they identified the three strongest correlates of reading comprehension: grammar knowledge (r = .85), vocabulary knowledge (r = .79), and decoding skills (r = .56). Here the strength of the correlation for vocabulary is numerically smaller than that of grammar, but it is important to note that this meta-analysis included studies with children and there was a statistically significant moderating effect

of age for vocabulary knowledge; the average correlation between vocabulary and reading was .84 for studies with adults (the age of 13 or older), whereas it was .66 for studies with children (the age of 12 or younger). Thus, for adult L2 learners, vocabulary knowledge seems to be at least as important as grammar knowledge for L2 reading comprehension.

To take a closer look at the role of linguistic knowledge components in L2 reading comprehension for the population of interest (i.e., adult learners), it may be useful to review individual studies with adult participants with multiple important variables (vocabulary, grammar, and decoding skills). For example, the study by Mecartty (2000) with adults with tests of vocabulary and grammar—the study mentioned in the preceding section on listening—actually examined reading comprehension skills as well. As with the listening test, the L2 Spanish learners' scores on the reading test were moderately positively correlated with their vocabulary and grammar knowledge (r = .50 and .34, respectively). However, here again, a multiple regression analysis showed that vocabulary knowledge was the only significant predictor when both vocabulary and grammar were entered into a model to predict reading comprehension. Therefore, this study suggests the importance of vocabulary knowledge in L2 reading comprehension as well (see also, e.g., Haynes & Carr, 1990; Khaldieh, 2001 for similar findings that vocabulary was somewhat more important than grammar for L2 reading comprehension).

On top of vocabulary and grammar knowledge, Nassaji and Geva (1999) examined two types of L2 decoding skills in L2 reading comprehension. One type—what they call 'phonological processing skill'—was the ability to convert spelling patterns into phonological codes and was assessed by a same/different decision task on pseudowords (e.g., *flemb-flem*). The other type—what they call 'orthographic processing skill'—was the knowledge of orthographic

regularity and was assessed by a binary-choice test where the participants selected pseudowords that they thought conformed to English orthographic conventions (e.g., *gmub*-*gnub*). Grammar knowledge was assessed by a grammaticality judgment test, and vocabulary knowledge was assessed by a multiple-choice test where the participants had to select the most appropriate definitions of target words. Sixty adult Iranian learners of English completed these tests as well as a reading comprehension test. The results showed that the participants' phonological and orthographic processing skills were positively correlated with their reading comprehension skills (r = .30 and .33, respectively). Their knowledge of vocabulary and grammar showed somewhat stronger correlations with their reading skills (r = .59 and .44, respectively).

In sum, overall vocabulary knowledge appears to be important not only for listening but also for reading comprehension.

1.7 Modality of Vocabulary Knowledge

Because of the importance of vocabulary knowledge for comprehension, there have been many studies that included the variable of vocabulary knowledge (see, e.g., Zhang & Zhang, 2022, who were able to include over 100 studies in their meta-analysis). However, one notable thing is that vocabulary knowledge tends to be measured in the written mode even when the primary focus of the study includes listening comprehension (e.g., Mecartty, 2000; Stæhr, 2008, 2009; Wang & Treffers-Daller, 2017). The meta-analysis by Zhang and Zhang (2022) in fact revealed that, whereas modality was consistent nearly 70% for the studies of reading comprehension (written vocabulary test, n = 72 vs. aural vocabulary test, n = 37), modality was consistent only about 30% for the studies of listening comprehension (aural vocabulary test, n = 15 vs. written vocabulary test, n = 30).

This inconsistency of modality in vocabulary tests is not only at odds with a multidimensional view of vocabulary knowledge (e.g., Nation, 2013), but also with a mainstream theory of lexical access. Bock and Levelt's (1994) model of lexical access, for instance, assumes three levels of representation: the conceptual level, the lemma level, and the lexeme level. At the conceptual level, a given entity is purely conceptual without any verbal information. For example, an entity, SHEEP, is an idea of the farm animal with thick wool. At the lemma level, a given entity has syntactic properties but without a manifested form. For example, an entity, SHEEP, is in the category of nouns (grammatical gender may also be marked for some languages). At the lexeme level, a given entity has form properties. For example, an entity, SHEEP, is expressed in a string of sounds /ʃip/ in the case of spoken language, and expressed in a string of letters *sheep* in the case of written language. Importantly, it is this lexeme level that sets aural vocabulary knowledge apart from written vocabulary knowledge. Without the consideration of modality, the validity of vocabulary tests could be diminished.

To address the scarcity of aural vocabulary tests, McLean et al. (2015), for example, developed the Listening Vocabulary Levels Test, an aural test of English vocabulary knowledge. This test covers 1000 to 5000 word-frequency levels as well as academic words with a total of 150 items. For each item test-takers listen to a target word followed by the word in a non-defining context sentence (e.g., *School: This is a big school*) and then select what they think is the most appropriate translation of the target word out of four options written in their L1. McLean et al. (2015) reported that scores on this test were positively correlated with scores on a listening comprehension test ($r = .54$) for a group of Japanese learners of English ($N = 214$).

In recent years the number of studies that include aural tests has been increasing (e.g., Cheng et al., 2023; Lange & Matthews, 2020; Masrai, 2022; Matthews & Cheng, 2015;

Matthews et al., 2024; Saito et al., 2023; Satori, 2021). These studies tend to show a strong correlation between an aural vocabulary test and a listening comprehension test (e.g., $r = .74$ in Masrai, 2022; $r = .94$ in Matthews et al., 2024). However, it is important to note that the relationship among variables is sample-dependent and therefore it is difficult to interpret results across studies. For instance, it is possible that a correlation turns out weak when the level of a test is not appropriate for a given sample (see, e.g., Lange & Matthews, 2020 for a weak correlation, $r = .15$, between vocabulary and listening presumably because the vocabulary test was too difficult for the participants). To examine the effect of modality (aural vs. written) more precisely then, both aural and written tests might need to be included in a single study.

Including both aural and written tests, Milton and Hopkins (2006) examined the effect of modality in vocabulary knowledge more closely. The researchers used parallel tests across modalities and so it was possible to compare aural vocabulary knowledge with the written equivalent. In particular, the researchers used two tests called X-Lex and Aural Lex; in X-Lex English words were presented one by one on the computer screen and the test-takers indicated (with yes or no) if they knew the meaning of the words; in Aural Lex English words were presented aurally with all else being equal. The study included two groups of L2 English learners, 88 L1 Greek learners and 38 L1 Arabic learners. The results showed that the Greek group scored significantly higher on the written test than on the aural test (out of 5000, $M = 2699$ for the written test vs. $M = 2017$ for the aural test). Meanwhile, the Arabic group scored higher on the aural test than on the written test ($M = 2554$ for the written test vs. $M = 2824$ for the aural test). A follow-up study by Milton et al. (2010) also reported that L1 Chinese learners showed an even greater difference between the written and aural tests ($M = 3272$ for the written test vs. $M = 2394$ for the aural test). These findings suggest that vocabulary knowledge is affected by the

mode of delivery and the effect may vary depending on the learners' L1 and/or their L2 learning backgrounds.

The effect of modality was reported with other tests and with another L1 group as well. Hamada and Yanagawa (2024), for example, had 155 Japanese learners of English take the Listening Vocabulary Levels Test (McLean et al., 2015) and its parallel written test. Each target word was presented first as a single word and then as a word in a non-defining context sentence either in the aural or written mode. The two tests had a common multiple-choice answer format where the learners selected one out of four written options as a L1 translation of the target word. The results showed that the Japanese learners scored significantly higher on the written test than on the aural test (out of 150, $M = 121$ for the written test vs. $M = 112$ for the aural test). Similarly, with a group of Japanese learners of English ($N = 332$), Mizumoto and Shimamoto (2008) examined the effect of modality in vocabulary knowledge. In their tests for each item a L1 word was first presented on paper and then four English words were presented. The learners were asked to choose an English word that corresponded to the L1 word. The only difference between their written and aural tests was the mode of delivery of the multiple-choice answer options. The researchers found a significant effect of modality and the learners' scores were higher for the written test than for the aural test (out of 160, $M = 118$ for the written test vs. $M = 110$ for the aural test).

Also, Uchihara (2023) examined the effect of aural versus written modality longitudinally by manipulating the modality of weekly quizzes in Japanese college EFL classes. Using a quasi-experimental design, the researcher administered weekly vocabulary quizzes in the aural mode in one class and in the written mode in another (equivalent) class over 10 weeks. The assumption here was that the students would study for the quizzes with the corresponding modalities. The

test format of the quizzes was meaning recall where the students provided L1 translation of L2 words presented either in the aural or written mode. Both groups also took pre- and post-tests of vocabulary in both modalities with the same meaning recall format. The results showed that the aural test group improved in aural vocabulary significantly more than the written test group, although there was no group difference in written vocabulary.

The studies reviewed in the preceding paragraphs showed that L2 learners often develop different levels of vocabulary knowledge in the aural and written modes. This inequality arises probably because learners tend to have different amounts of instruction and practice across the modalities and also because there are modality-specific challenges in form-meaning mapping. These challenges have been documented by psycholinguistic studies. Cutler et al. (2006), for instance, demonstrated Japanese learners' difficulty perceiving English /l/ and /r/ phonemes using an eye-tracking paradigm. They found that when Japanese learners were told to click on a picture of a rocket, for example, they were distracted and looked at a picture of a locker if both pictures were present. This finding suggested that the participants' form-meaning mapping was incomplete in the aural mode. Similarly, Pallier et al. (2001) demonstrated Spanish speakers' difficulty perceiving Catalan-specific phonemic contrasts using a priming paradigm. They found that Spanish-Catalan bilinguals showed repetition priming for Catalan minimal pairs (e.g., /netə/, *granddaughter*, and /nɛtə/, *clean*) in an aural lexical decision task when the participants' native language was Spanish rather than Catalan, suggesting that the native Spanish speakers were treating the Catalan minimal pairs as the same words. Thus, although these native Spanish speakers were fluent in Catalan, their Catalan form-meaning mapping was not fully functional in the aural mode.

Considering the facts that L2 learners could have different levels of vocabulary knowledge in the aural and written modes and that there are modality-specific challenges in form-meaning mapping, it may be hypothesized that L2 vocabulary knowledge consists of aural and written vocabulary. Alternatively, it is also possible that there is general vocabulary knowledge and the knowledge is influenced by modality-specific skills (e.g., phonological/orthographic decoding abilities).

1.8 Aural and Written Vocabulary and L2 Comprehension

If there is a distinction between aural and written vocabulary knowledge, it would be expected that the knowledge drawn upon during listening comprehension is aural vocabulary rather than written vocabulary (and the reverse is true for reading comprehension). Therefore, a stronger association would be expected between aural vocabulary and listening skills than between written vocabulary and listening skills. Similarly, a stronger association would be expected between written vocabulary and reading skills than between aural vocabulary and reading skills. The findings of a meta-analysis by Zhang and Zhang (2022) are consistent with these predictions. The researchers included 126 correlational studies and found that listening comprehension skills were on average correlated more strongly with aural vocabulary (r = .60) than with written vocabulary (r = .52), whereas reading comprehension skills were correlated more strongly with written vocabulary (r = .60) than with aural vocabulary (r = .49). These findings appear to suggest the importance of modality of vocabulary knowledge in L2 listening and reading comprehension.

To take a closer look at the relationships between aural and written vocabulary knowledge and listening and reading comprehension with greater comparability, studies with

adult learners that included *both* aural and written vocabulary tests were examined. The studies are summarized in Table 2. In most of the studies, listening comprehension test scores correlated more strongly with aural vocabulary test scores than with the written counterparts, and reading comprehension test scores correlated more strongly with written vocabulary test scores than with the aural counterparts. These observations are consistent with the findings of the meta-analysis by Zhang and Zhang (2022) mentioned above.

Table 2

Summary of Correlations Among Aural Vocabulary, Written Vocabulary, Listening Comprehension, and Reading Comprehension

Study	L1	A/W vocab tests parallel format?	A/W vocab tests correlation	Aural vocab test & listening correlation	Written vocab test & listening correlation	Aural vocab test & reading correlation	Written vocab test & reading correlation
Cheng & Matthews (2018)	Chinese	No	(1) .59* (2) .72*	**.71***	(1) .39* (2) .55*	.46*	(1) .46* (2) **.57***
Ha (2021)	Vietnamese	No	.89*	.65*	.65*	.61*	**.62***
Hamada & Yanagawa (2024)	Japanese	Yes	.70*	.34*	**.47***	N/A	N/A
Irvine et al. (1974)	Persian	No	.47*	**.69***	.56*	**.53***	.49*
Masrai (2020)	Various	No	.58*	**.67***	.59*	N/A	N/A
Milton & Hopkins (2006)	Greek & Arabic	Yes	.68*	N/A	N/A	N/A	N/A
Milton et al. (2010)	Various	Yes	.46*	**.67***	.48*	.22	**.70***
Mizumoto & Shimamoto (2008)	Japanese	Yes	.89*	**.60***	.56*	.69*	**.72***
Uchihara & Harada (2018)	Japanese	No	.73*	N/A	N/A	N/A	N/A

Note. *$p < .05$; A = aural; W = written. Cheng and Matthews (2018) included two written vocabulary tests and so they are marked as (1) and (2). Correlation coefficients in bold with grey shading indicate the stronger correlation (aural vs. written vocabulary) with comprehension (listening or reading) within each study. The target language was English for all the studies.

Finally, a study by Cheng and Matthews (2018) provided additional information, going beyond simple correlational analysis using hierarchical multiple regression and exploratory factor analysis. The researchers had 250 Chinese learners of English take one aural and two written vocabulary tests as well as listening and reading comprehension tests. One written vocabulary test was a receptive test where the learners matched target words with their meanings, and another written test was a productive test where the learners wrote target words prompted by their initial letters and contextual sentences. The aural vocabulary test was partial dictation where the learners filled out target words after listening to sentences. The results of hierarchical regression analysis showed that the scores on the aural vocabulary test explained about 51% of variance in listening comprehension skills and the addition of the two written vocabulary tests as predictors improved the model only by 1%. For reading comprehension skills, the scores on the written productive test explained about 33% of the variance and the addition of the other two predictors did not improve the model. Furthermore, the researchers conducted factor analysis on the vocabulary tests. They found two factors with eigenvalues greater than one. One factor was loaded by the two written tests, and another factor was loaded by the aural test. Therefore, this study provides further evidence for the significance of modality in vocabulary knowledge and in comprehension skills.

1.9 A Gap in the Literature

Although the existing body of research may be able to support the importance of modality overall, those studies also have limitations. One limitation pertains to measurement error. As widely recognized, a correlation coefficient is attenuated when measures are unreliable, and so "observed correlation is a lower bound of true correlation" (Raykov & Marcoulides, 2011, p. 192). In other words, lack of reliability of measures might have been obscuring the true

relationship between the variables of interest. Another limitation in the literature concerns spurious association between variables due to (dis)similarity of measures. For example, when aural and written vocabulary knowledge is measured with the same test format (e.g., multiple-choice) and with the same target words (as in, e.g., Mizumoto & Shimamoto, 2008), correlation between the variables would be inflated by the similarities of the measures that are outside of the construct of interest. Conversely, when aural and written vocabulary knowledge is measured with very different tests, for instance, with dictation for the former and with multiple-choice recognition test for the latter (as in, e.g., Cheng & Matthews, 2018), correlation between the variables would be deflated by the dissimilarities of the measures that are outside of the construct of interest. These limitations—measurement error and spurious association—may not be overcome as long as the construct of interest is assessed at the level of measured variables.

Structural equation modeling has a potential to overcome the limitations. Using multiple measures to capture a single latent construct allows for removing systematic characteristics of measures from the construct under investigation. Measurement error can also be accounted for with this approach. Moreover, this technique allows for testing hypothesized structural relations derived from theory. For example, it is possible to model the latent constructs of general vocabulary knowledge and aural and written modality-specific skills separately and evaluate how each of the constructs relates to listening and reading comprehension skills. To my knowledge, this type of modeling has not been done in this line of research, and it may help better understand the significance of modality in vocabulary knowledge and in comprehension skills.

1.10 Previous Studies of the Dimensionality of Vocabulary Knowledge With Structural Equation Modeling

Although the effect of aural versus written modality does not seem to have been examined with SEM, several studies have examined other aspects of vocabulary knowledge with SEM. It might be useful to review these studies before diving into the current investigation.

González-Fernández and Schmitt (2020) examined the dimensionality of vocabulary knowledge in terms of breadth and depth for Spanish learners of English (N = 144). They included four vocabulary knowledge components: the form-meaning link, derivatives, multiple meanings, and collocations. Each of the components was assessed by receptive (recognition) and productive (recall) tests. Initially, the researchers developed a second-order factor model of vocabulary knowledge, where the factor of general vocabulary knowledge influenced the factors of four vocabulary knowledge components with each of the four factors indicated by two (receptive and productive) test scores. Although the data-model fit of this model was overall acceptable, the standardized factor loadings from the general vocabulary factor to the four knowledge component factors were very high (.94 to .98). The researchers considered these loadings too high and developed an alternative, one-factor model. The one-factor model showed a better data-model fit, suggesting unidimensionality of vocabulary knowledge. This finding was replicated with L1 Chinese learners of English in González-Fernández (2022).

Similarly, Koizumi and In'nami (2020) examined the dimensionality of vocabulary knowledge in terms of breadth and depth for Japanese learners of English (N = 255). They included tests of vocabulary size, word association, polysemy, and collocation, and compared a one-factor model and a two-factor model. The two-factor model consisted of a breadth factor indicated by subtests of vocabulary size, and a depth factor indicated by tests of word

association, polysemy, and collocation. While the data-model fits of both models were acceptable, a chi-square difference test between the two models was statistically significant, indicating that the two-factor model had a significantly better fit to the data than the one-factor model. However, the breadth and depth factors in the two-factor model were highly correlated (r = 0.95), suggesting a considerable overlap between the two factors.

Stewart et al. (2024) examined the dimensionality of vocabulary knowledge in terms of test type (recognition vs. recall) for Japanese learners of English (N = 103). They included two types of vocabulary tests: a meaning recognition test where the participants selected an appropriate L1 translation out of four options for each target word, and a meaning recall test where the participants translated each target word into L1. The researchers compared a one-factor model and a two-factor model. The two factors in the latter model were indicated by subsets of the recognition and recall test items, respectively. The results showed that, whereas the one-factor model did not fit the data well, the two-factor model demonstrated an acceptable data-model fit, suggesting that there is a difference between recognition and recall tests. That said, the factor correlation was high (r = 0.85), indicating a large overlap between the two factors.

Hui et al. (2022) examined the dimensionality of vocabulary knowledge in terms of availability of time in test response for L2 learners of English (N = 145). They included two kinds of vocabulary tests, one in which the participants could spend as much time as they liked, and another in which the participants had to respond as quickly as possible. The researchers compared a one-factor model and a two-factor model. In the two-factor model one factor was indicated by scores on untimed tests of meaning recognition and form recall and the other factor was indicated by scores on time-sensitive tests of yes-no decision on form-meaning knowledge

(indexed by accuracy and reaction time) and lexical decision (indexed by repetition priming effect). The results showed that both one-factor and two-factor models had acceptable data-model fits, and that a chi-square difference test between the two models was non-significant, meaning that incorporating this time factor into the model did not improve data-model fit. Furthermore, the factor correlation in the two-factor model was high ($r = 0.92$), indicating a considerable overlap between the two factors. Therefore, this study suggested unidimensionality of vocabulary knowledge with regard to the aspect of time-sensitiveness.

In sum, some studies showed some aspects of multidimensionality of vocabulary knowledge (Koizumi & In'nami, 2020; Stewart et al., 2024), while others did not show multidimensionality (González-Fernández, 2022; González-Fernández & Schmitt, 2020; Hui et al., 2022). Interestingly, however, in every case, factor correlation in a multi-factor model was high, indicating somewhat limited impact of additional dimensions beyond general vocabulary knowledge. Whether a similar result will be obtained for the dimension of aural versus written modality or not is an unanswered question.

1.11 Present Study

In this study I examined the extent to which L2 vocabulary knowledge is domain-general or modality-specific, that is, how much L2 aural and written vocabulary is separable and explains L2 listening and reading comprehension differently. Adult L2 learners in a foreign language context took four aural and four written vocabulary tests, with parallel test formats across the modalities to allow for comparison. The participants also took listening and reading comprehension tests. The data was analyzed by SEM. To my knowledge, this study was the first attempt to tease apart modality-specific skills from L2 vocabulary knowledge at the level of

latent constructs, thereby having the potential to clarify the role of modality in vocabulary and comprehension with less contamination by measurement characteristics than previous studies.

First, in this study the structure of L2 vocabulary knowledge was examined:

RQ1: Is adult L2 learners' vocabulary knowledge unidimensional or bi-dimensional in terms of the aural and written modalities?

To answer this question, three possible vocabulary knowledge structures were formed and compared: (1) one-factor structure with general vocabulary (no modality effect), (2) two-factor structure with aural and written vocabulary, and (3) residualized-factor structure with general vocabulary and aural and written modality-specific skills (see the Analysis section for the details).

Prediction: Based on the previous studies that suggested modality effects in L2 vocabulary tests (e.g., Cheng & Matthew, 2018; Milton & Hopkins, 2006), I predicted that L2 vocabulary knowledge would be bi-dimensional. Given the lack of empirical evidence, I did not have any specific predictions about which of the two bi-dimensional structures would show a better data-model fit.

Second, using latent variable path analysis, it was examined whether the modality-specific aspects of vocabulary knowledge explain L2 listening and reading comprehension skills:

RQ2a: To what extent does adult L2 learners' aural vocabulary knowledge uniquely explain their listening comprehension skills?

RQ2b: To what extent does adult L2 learners' written vocabulary knowledge uniquely explain their reading comprehension skills?

Prediction: I predicted that aural and written vocabulary knowledge would uniquely explain listening and reading comprehension skills, respectively. This prediction was based on the higher correlations of aural and written vocabulary with comprehension skills in the respective domains in previous studies (e.g., Cheng & Matthew, 2018; Milton et al., 2010; Zhang & Zhang, 2022).

Third, if the best model of vocabulary knowledge structure consisted of two components, aural and written vocabulary, the size of vocabulary would be compared. Using latent means modeling, it would be examined whether aural vocabulary was larger, smaller, or equivalent to written vocabulary for the current group of participants:

RQ3: Is adult L2 learners' aural vocabulary larger, smaller, or equivalent to their written vocabulary?

Prediction: Previous studies suggested that adult L2 learners in a foreign language context might have smaller aural vocabulary than written vocabulary (e.g., Hamada & Yanagawa, 2024; Mizumoto & Shimamoto, 2008), and therefore I predicted that aural vocabulary would be smaller than written vocabulary for the current participants.

In this study, a sample was drawn from L1 Japanese learners of English in Japan. These L2 learners tend to have a learning background with greater focus on written than aural language partly because they often need(ed) to pass high-stakes written English tests for college admissions (Kobayashi, 2001). Given this learning background, they were a good target population for examining the differences between aural and written vocabulary as they were expected to have unequal sizes of aural and written vocabulary (i.e., larger written vocabulary than aural vocabulary). This also means that it is important to keep in mind limited generalizability of the current findings. Other L1 groups, for example, might show different

relative levels of aural and written vocabulary knowledge. In addition, child L2 learners, immigrant learners, or even more extreme—non-literate learners—would show a different level and/or structure of vocabulary knowledge. Therefore, the results of the present study should not be generalized to different populations without replication.

Chapter 2: Methodology

In the spirit of open science, the research design and analysis methods were pre-registered prior to data collection and are publicly available in the Open Science Framework at https://osf.io/rkc4b.

2.1 Participants

Adult learners of English were recruited in Japan. The participants needed to meet all the following eligibility criteria: (1) between the ages of 18 to 40, (2) native speaker of Japanese, (3) no substantial exposure to English before the age of 18 (i.e., growing up in a monolingual family and having education in L1), and (4) no experience staying abroad for more than a month. The age and L1 criteria were included to minimize confounds related to aging and L1 influence, respectively. The latter two criteria were included to specifically target the population of interest; a sample had to be drawn from late L2 learners with limited overseas experience because in this study I was interested in how modality (aural and written) played a role among adult learners in a foreign language context (see, e.g., DeKeyser & Larson-Hall, 2005 for a discussion of the critical period and how learning process could be different between children and adults).

As in Hui et al. (2022), sample size planning was conducted (see also, e.g., Loewen & Hui, 2021 for a discussion of sample size planning). A minimum sample size of 160 was determined for this study considering the current research purpose and analysis framework as well as the sample sizes in similar previous studies. First, the primary research question of this study (i.e., dimensionality of vocabulary knowledge) rested on the evaluation of data-model fit. The most complex SEM model of vocabulary knowledge had eight degrees of freedom (see the Analysis section), and based on previous studies (e.g., González-Fernández, 2022; Hui et al.,

2022), expected factor loadings were around .80. According to Kim (2005), for a model with eight degrees of freedom, average factor loadings at .80, and Comparative Fit Index (CFI) = .95 to achieve a power of .80 (in terms of the assessment of global data-model fit), the proposed minimum sample size was 127. Also, for previous studies on the dimensionality of vocabulary knowledge with SEM, González-Fernández and Schmitt (2020) had 144 participants, González-Fernández (2022) had 170, and Hui et al. (2022) had 145. Therefore, around 150 seemed to be a common sample size in previous studies. Given these considerations (and potential exclusion of some portion of data), 160 was considered reasonable for this study.

The participants were recruited by posting flyers on social media platforms, websites, listservs, and so forth (i.e., convenience sampling). They received financial compensation, a JP¥4000 gift card (about 28 USD), at the end of the study. Following the guidelines of the Institutional Review Board, the participants completed a consent form before the study.

In the end, 185 people participated in this study. There were 108 females (58%) and 77 males (42%). The average age of the participants was 23.02 years (SD = 4.28, range = (18, 39), median = 22, mode = 21). The majority of the participants were undergraduate students (n = 125; 68%). The rest of the participants were young professionals (n = 37; 20%), graduate students (n = 20; 11%), unemployed (n = 2; 1%), and a housewife (n = 1; 0.5%). The participants had a diverse background of college majors. The top five most frequent majors were economics (n = 19; 10%), law (n = 16; 9%), computer science (n = 13; 7%), English (n = 13; 7%), and Japanese (n = 13; 7%). See Appendix A for a comprehensive list of the participants' college majors.

2.2 Vocabulary Tests

2.2.1 Overall Vocabulary Test Design

There were eight measures of vocabulary knowledge, four written and four aural vocabulary tests. All the tests measured vocabulary size in terms of form-meaning mapping. There were four test formats: form recall, meaning recall, yes-no decision, and meaning selection. These formats were used across modalities; therefore, written and aural tests had the same test formats, making it possible to compare the two modalities.

Different sets of target words were used across different vocabulary tests along with counterbalancing. This choice was made in light of the primary goal of the present study, that is, to compare model fit between one-factor and multi-factor models. For fair model comparison in SEM, careful thought needs to be given to the selection of indicators, bearing in mind what the target latent construct is and what the method is through which the construct is examined. The following quote from Hancock and Schoonen (2015) may help make the choice of indicators in SEM:

> In addition to issues of reliability, validity of factors vis-à-vis their indicators can be a challenge as well, in particular when indicators are all from the same type of data source. Using all retrospective questionnaires as indicators, for example, however reliable they might be, can create factors that confound the latent traits of interest with the methods of measurement. Whenever possible, then, indicators should be chosen from multiple modes of measurement (e.g., survey, behavioral observation) to make sure that each factor separates trait from method as much as possible. (p. 179)

For the present study, the obvious implication of this quote was that various formats of vocabulary tests (multiple choice, recall, etc.) should be used for each construct. What to do with target words was a somewhat more subtle issue and required careful consideration. In the case of the present study, the latent construct of interest was learners' *level* of vocabulary (whether it was written vocabulary, aural vocabulary, or general vocabulary), rather than the knowledge of any particular words. Therefore, target words were a sample of words with which learners' level of vocabulary was probed. If the same set of words had been used across all indicators (as in previous studies, e.g., González-Fernández, 2022; Hui et al., 2022), the unique characteristics of the words—which was an aspect of method—would have contributed to the overall commonality and potentially biased the data for the one-factor model. Some might think that using different sets of words, in turn, would bias the data in favor of the multi-factor model, but that would not be the case because, when a different set of words is used for each indicator, it does not help form any of the factors—the unique characteristics of words would simply become error variance without favoring either of the models. Thus, it should be fair in terms of model comparison. For this reason, different sets of words were used across indicators in this study.

2.2.2 Target Words

The purpose of the current vocabulary tests was to estimate the participants' vocabulary size (level) relevant to listening and reading comprehension. To this end, an adequate sample of vocabulary needed to be selected. The target words were chosen from 2000 to 5000 frequency level. This decision was made based on the findings that (1) vocabulary knowledge up to and including the 5000 frequency level seemed to be necessary for adequate reading and listening comprehension (Adolphs & Schmitt, 2003; Laufer & Ravenhorst-Kalovski, 2010), and that (2) the participants similar to those in the present study reached a ceiling score for 1000 frequency

level words (see, e.g., Mizumoto & Shimamoto, 2008). The JACET8000 vocabulary list (JACET Basic Word Revision Committee, 2016) was used for selecting the target words (for other studies with this list, see, e.g., Koizumi & In'nami, 2020; Mizumoto & Shimamoto, 2008). This vocabulary list was made based on the British National Corpus and the Corpus of Contemporary American English, while also taking into account local materials that the participants in this study (i.e., people in Japan) were likely to encounter (e.g., textbooks). The proportion of the parts of speech of the target words in the present study was: 50% nouns, 30% adjectives, and 20% verbs, which was similar to the ratio adopted by Cheng and Matthews (2018) (50% nouns, 25% adjectives, 25% verbs) and roughly reflected the proportion of the JACET8000 vocabulary list (57% nouns, 21% adjectives, 13% verbs).

Eight sets of 40 target words were created for the eight vocabulary tests. Each set consisted of 10 words (5 nouns, 3 adjectives, and 2 verbs) from each of the four frequency bands from 2000 to 5000 frequency level. The procedure for creating the sets was as follows: Forty nouns, 24 adjectives, and 16 verbs were randomly selected from each of the four frequency bands. Then the selected words were examined and cognates (e.g., *toilet* in English vs. *toire* in Japanese) were replaced with noncognates (which were newly randomly selected from the same frequency band) because cognates would be too easy. After that, eight subsets of 10 words were created for each frequency band with the same part-of-speech ratio. These eight subsets were made such that they were roughly equivalent to one another in terms of the number of phonemes, the number of letters, and frequency. The number of phonemes was calculated with the Phoneme Counter Ver5.1 (Nakanishi, 2019). Then, the subsets were combined such that each of the eight sets had one subset from each of the four frequency bands. Since the vocabulary tests included translation between L1 and L2, words with unclear translation were not desirable; therefore,

three English-Japanese dictionaries were consulted and the words for which the first entry (primary meaning) was different in every dictionary were replaced. Also, some words were too short to construct test prompts and therefore they were replaced. The final eight sets of target words are found in Table 3. There was no significant difference across the eight sets in the number of phonemes, $F(7, 312) = 0.02$, $p = 1$, the number of letters, $F(7, 312) = 0.01$, $p = 1$, and frequency, $F(7, 312) = 0.01$, $p = 1$. Furthermore, the sets of target words were examined in terms of difficulty (how much a given word is known by L2 speakers). Difficulty information was drawn from a large-scale study by Brysbaert et al. (2021), who reported accuracy rates of 61851 English words in a recognition test based on 17 million responses by L2 speakers. Using the accuracy rates as an index, there was no significant difference in difficulty across the eight sets of target words, $F(7, 312) = 0.73$, $p = .647$.

Table 3

Sets of Target Words

Set A	Set B	Set C	Set D	Set E	Set F	Set G	Set H
			2000-word frequency				
ocean	hobby	fever	exam	rent	detail	habit	finger
honor	poem	salt	nation	umbrella	sugar	award	sand
enemy	monkey	income	trust	illness	temple	emotion	cousin
interest	result	citizen	ability	facility	diary	belief	origin
ceiling	direction	imagination	factor	object	homework	effect	pencil
solar	tiny	polite	ideal	sunny	alive	recent	female
usual	chemical	honest	accurate	obvious	gentle	personal	narrow
secret	monthly	delicious	stable	historical	elderly	cultural	particular
employ	exist	divide	whisper	suggest	attach	expect	escape
borrow	ignore	obtain	attend	rely	fold	hurry	absorb
			3000-word frequency				
pepper	shelf	prison	arrival	weapon	wealth	wound	poison
pond	nest	crisis	editor	labor	aspect	author	honey
depth	wisdom	length	onion	treasure	harvest	liquid	basis
anger	departure	emphasis	quantity	insurance	valley	mosquito	candidate
fossil	investment	relation	election	exit	protest	phenomenon	psychology
proper	solid	vast	vital	urban	legal	guilty	noisy
absent	regional	curious	critical	grateful	annual	urgent	generous
desirable	sudden	awake	entire	eager	fragile	nuclear	absolute
deny	persuade	donate	occur	declare	admire	educate	melt
pronounce	possess	hesitate	calculate	postpone	punish	explode	eliminate
			4000-word frequency				
needle	feather	occupation	equality	insult	laundry	obligation	voyage
organ	burden	mammal	divorce	legend	puppy	bamboo	pupil
patience	lottery	wallet	decay	heaven	surgeon	ritual	ladder
currency	tuition	gravity	justice	servant	envelope	founder	funeral

frequency	luxury	volcano	contractor	climber	grammar	margin	dialect
vertical	windy	polar	naked	bitter	tragic	visible	tense
fluent	extinct	miserable	genetic	ultimate	occasional	oral	durable
deliberate	passive	gradual	lazy	profitable	conservative	intensive	pregnant
vanish	pollute	obey	forgive	renew	emit	insert	invade
evolve	expire	distort	strengthen	exploit	stimulate	inherit	exceed
5000-word frequency							
wolf	riot	arrow	asset	horizon	bulb	saint	voter
commuter	rumor	copper	erosion	vinegar	virtue	theft	chaos
colony	dignity	liberty	axis	worship	suicide	harbor	apology
completion	hypothesis	odor	aviation	destiny	subsidy	prevention	patent
hostage	contributor	manuscript	murder	bully	coincidence	fatigue	prescription
immune	hopeful	moist	swift	muddy	rigid	cruel	racial
decisive	shallow	cellular	elaborate	bold	supreme	optimistic	mutual
continental	vacant	decent	infinite	municipal	dietary	tidy	mandatory
notify	revive	embed	omit	tilt	collide	resign	worsen
tremble	deprive	violate	exaggerate	cooperate	overlook	embrace	conceal
Average number of phonemes in the set (*SD*)							
5.53 (1.65)	5.55 (1.63)	5.60 (1.68)	5.60 (1.48)	5.60 (1.52)	5.55 (1.57)	5.60 (1.60)	5.60 (1.68)
Average number of letters in the set (*SD*)							
6.58 (1.74)	6.58 (1.66)	6.53 (1.68)	6.58 (1.71)	6.58 (1.66)	6.60 (1.71)	6.58 (1.57)	6.58 (1.71)
Average word frequency in the set (*SD*)							
2989 (1183)	3014 (1161)	3008 (1272)	2999 (1208)	3025 (1135)	3021 (1192)	3013 (1207)	2986 (1209)

2.2.3 Written and Aural Vocabulary Test Formats

Four test formats were included in this study because at least four indicators were required for each vocabulary factor so that all the models of vocabulary knowledge structure would be over-identified models (i.e., statistically testable models where the number of parameters to be estimated is fewer than the number of unique variances and covariances). To minimize method effects, it was preferable to include various formats of tests, while at the same time these formats should be representative of the tests in the field of vocabulary research. According to a meta-analysis by Zhang and Zhang (2022), there are mainly three formats of tests in previous studies of L2 vocabulary: form recall, meaning recall, and meaning recognition. As the first two formats are productive tests and the third one is a receptive test, it was decided that the present study would include one form recall test, one meaning recall test, and two meaning recognition tests, thereby consisting of two productive and two receptive tests. Yes-no decision and meaning selection were selected for the meaning recognition tests, both of which have been frequently used in the studies of L2 vocabulary (for yes-no decision, see, e.g., Brysbaert et al., 2021; Hui & Godfroid, 2021; Hui et al., 2022; Masrai, 2020, 2022; McLean et al., 2020; Meara, 2010; Milton & Hopkins, 2006; Milton et al., 2010; Mochida & Harrington, 2006; and for meaning selection, see, e.g., Andringa et al., 2012; González-Fernández, 2022; González-Fernández & Schmitt, 2020; Ha, 2021; Hamada & Yanagawa, 2024; Henning et al., 1981; Hui et al., 2022; Lange & Matthews, 2020; McLean et al., 2015, 2020; Mecartty, 2000; Nassaji & Geva, 1999; Saito et al., 2023; Satori, 2021; Stewart et al., 2024; Uchihara & Harada, 2018; Vafaee & Suzuki, 2020; Wallace, 2022; Wang & Treffers-Daller, 2017; Wong et al., 2017). Each of the test formats will be explained in detail in the following subsections. It would be useful to note that the four test formats were assumed to be different in terms of difficulty too. For

example, a form recall test, where participants had to produce target words by themselves, was expected to be more difficult than a meaning selection test, where they only needed to select meaning of target words. This wide range of difficulty levels was important for the present study as the study included participants with a wide range of proficiency levels.

It may also be important to note that the present study examined each of written and aural vocabulary knowledge as a whole, that is to say, the study viewed knowledge of form-meaning mapping in each modality as a set of skills and did not try to locate where problems occurred within the set of skills. For instance, when a participant failed to answer an aural meaning recognition test item correctly, it might be because the participant was not able to perceive phonemes correctly or it might be because the participant had imprecise lexical representations in their mental lexicon (among other possible reasons). It was not the scope of this study to identify such exact causes (but for a review of psycholinguistic literature on L2 learners' problems with phonetic perception and lexical representations, see, e.g., Gor, 2015).

Given the inherent differences between written and aural modalities, it was not possible to create exactly the same experimental condition across the modalities, but an attempt was made to make the presentation of written and aural stimuli comparable. Namely, the presentation time of target words was limited for both modalities: written target words were presented only for 3 seconds and aural target words were presented only once. For the form recall tests, initial letter/phoneme prompts were available to the participants throughout the test trials in both written and aural tests.

2.2.3.1 Form Recall Test of Written Vocabulary

This test measured the participants' ability to produce L2 words in the written mode when the L1 translation was given. In each item a L1 word was presented on the computer screen and the participants were asked to type the corresponding L2 word. The part of speech of the target word was also presented along with the L1 word in the participants' L1. To elicit target words, the initial letter(s) were also provided. There were 40 items in total. An example item is as follows:

Example question: 知恵 (名詞) wi_____

Example answer: wisdom

Translation of example question: wisdom (noun) wi_____

For previous studies with similar test formats, see, for example, Cheng and Matthews (2018), González-Fernández (2022), González-Fernández and Schmitt (2020), Hui et al. (2022), and McLean et al. (2020).

In the form recall tests, initial letter(s)/phoneme(s) were provided to prevent the participants from avoiding the target words. These prompts helped reduce the number of cases where the participants provided other acceptable words and maintain the validity of the tests as measures of vocabulary size (for previous studies with initial letter prompts, see, e.g., Cheng & Matthews, 2018; González-Fernández, 2022; González-Fernández & Schmitt, 2020; Hui et al., 2022). This task setup in turn, however, limited ecological validity in the sense that the

participants could not freely use their own language resources. It also means that the task was not form recall from scratch. These aspects should be acknowledged as limitations.

2.2.3.2 Form Recall Test of Aural Vocabulary

This test measured the participants' ability to produce L2 words in the aural mode when the L1 translation was given. The test was the same as the form recall test of written vocabulary, except for the following: the participants were asked to say (but not type) L2 words, and the initial phoneme(s) (but not letters) were provided. The initial phoneme(s) were played by the participants clicking a button on screen. There were 40 items in total. An example item is as follows:

Example question: 知恵 (名詞) /wɪ/

Example answer: /wɪzdəm/

Translation of example question: wisdom (noun) /wɪ/

There does not seem to be any previous study with this test format with adult participants (but see, e.g., Lesaux et al., 2010; Proctor et al., 2005 for previous studies with an aural form recall test in the form of picture naming for child participants). It appears that dictation has been used for a form recall test of aural vocabulary for adult learners (see, e.g., Bonk, 2000; Cheng & Matthews, 2018; Masrai, 2022; Matthews & Cheng; 2015).

2.2.3.3 Meaning Recall Test of Written Vocabulary

This test measured the participants' ability to produce L1 meaning when L2 words were given in the written mode. In each item a L2 word was presented on the computer screen for 3

seconds. The participants were then asked to type the meaning in their L1. The part of speech was also presented in their L1 when they responded. There were 40 items in total. An example item is as follows:

Example question: wisdom (名詞)

Example answer: 知恵

Translation of example question: wisdom (noun)
Translation of example answer: wisdom

For previous studies with similar test formats, see, for example, Jeon (2011), Khaldieh (2001), McLean et al. (2020), Stewart et al. (2024), and Uchihara (2023).

2.2.3.4 Meaning Recall Test of Aural Vocabulary

This test measured the participants' ability to produce L1 meaning when L2 words were given in the aural mode. The test was the same as the meaning recall test of written vocabulary, except for the following: L2 words were presented aurally but not visually. The audio was played once. There were 40 items in total. An example item is as follows:

Example question: /wɪzdəm/ (名詞)

Example answer: 知恵

Translation of example question: /wɪzdəm/ (noun)

Translation of example answer: wisdom

For previous studies with similar test formats, see, for example, Cheng et al. (2023), Matthews et al. (2024), and Uchihara (2023).

2.2.3.5 Yes-No Test of Written Vocabulary

This test measured the participants' ability to recognize L2 words when they were presented in the written mode. In each item a L2 word was presented on the computer screen for 3 seconds. The participants were then asked to select "Yes" if they knew the meaning of the word and, "No" if not. To prevent mindless "Yes" response, 20 nonword items were also included. This ratio (40 real words and 20 nonwords) was the same as in the original yes-no test (Meara, 2010). There were 60 items in total. An example item is as follows:

Example question: wisdom

Example answer: Yes

For previous studies with similar test formats, see, for example, Brysbaert et al. (2021), Hui et al. (2022), Masrai (2020), McLean et al. (2020), Milton and Hopkins (2006), Milton et al. (2010), and Mochida and Harrington (2006).

2.2.3.6 Yes-No Test of Aural Vocabulary

This test measured the participants' ability to recognize L2 words when they were presented in the aural mode. The test was the same as the yes-no test of written vocabulary, except for the following: L2 words and nonwords were presented aurally but not visually. The audio was played once. There were 60 items in total. An example item is as follows:

Example question: /wɪzdəm/

Example answer: Yes

For previous studies with similar test formats, see, for example, Hui and Godfroid (2021), Masrai (2020, 2022), Milton and Hopkins (2006), and Milton et al. (2010).

2.2.3.7 Meaning Selection Test of Written Vocabulary

This test measured the participants' ability to recognize L1 meaning when L2 words were presented in the written mode. In each item a L2 word was presented on the computer screen for 3 seconds. The participants were then asked to select the corresponding L1 word out of four options presented on screen. Distractors were L1 translation of English words randomly sampled from the same frequency band and the same part of speech as the given target word. There were 40 items in total. An example item is as follows:

Example question: wisdom

Example answer options: (1) 目的 (2) 観光 (3) 知恵 (4) 望遠鏡

Example answer: (3) 知恵

Translation of example answer options: (1) objective (2) sightseeing (3) wisdom (4) telescope

Translation of example answer: (3) wisdom

For previous studies with similar test formats, see, for example, Andringa et al. (2012), González-Fernández (2022), González-Fernández and Schmitt (2020), Hamada and Yanagawa (2024), Henning et al. (1981), Hui et al. (2022), McLean et al. (2020), Mecartty (2000), Nassaji and Geva (1999), Stewart et al. (2024), and Wang and Treffers-Daller (2017).

2.2.3.8 Meaning Selection Test of Aural Vocabulary

This test measured the participants' ability to recognize L1 meaning when L2 words were presented in the aural mode. The test was the same as the meaning selection test of written vocabulary, except for the following: L2 words were presented aurally but not visually. The audio was played once. There were 40 items in total. An example item is as follows:

Example question: /wɪzdəm/

Example answer options: (1) 準備 (2) 広告 (3) 知恵 (4) 動物園

Example answer: (3) 知恵

Translation of example answer options: (1) preparation (2) advertisement (3) wisdom (4) zoo

Translation of example answer: (3) wisdom

For previous studies with similar test formats, see, for example, Ha (2021), Hamada and Yanagawa (2024), Lange and Matthews (2020), McLean et al. (2015), Saito et al. (2023), Satori (2021), Uchihara and Harada (2018), Vafaee and Suzuki (2020), and Wallace (2022).

2.2.4 Vocabulary Test Material Development

For the yes-no tests, two sets of 20 nonwords were created, using the ARC Nonword Database (Rastle et al., 2002). Nonwords were selected from those with "orthographically existing onsets," "orthographically existing bodies," and "legal bigrams." After consultation with four native English speakers, nonwords that sounded too close to a real word were replaced. The nonword sets were similar to the sets of target words in terms of the number of phonemes and the number of letters and the two nonword sets were roughly parallel to each other. The sets of nonwords are found in Table 4. There was no significant difference between the two nonword sets or between the real and nonword sets in the number of phonemes, $F(9, 350) = 0.57$, $p = .821$, and the number of letters, $F(9, 350) = 0.16$, $p = .998$.

All audio materials (target words and nonwords) for the aural vocabulary tests were recorded by a male native speaker of American English. He was 22 years old at the time of recording and had lived in the state of Maryland for his entire life. The recording was conducted in a professional studio. The speaker was instructed to say each word with natural speed and with normal intonation. After the recording, the sound files were edited and normalized using Audacity. Phoneme prompts for the aural form recall test were created by carefully clipping the initial part of the sound files of the target words in Audacity. Phonemes taken from the target words were used as prompts as opposed to isolated phonemes because phonemes within a word are unique to the particular word due to various factors such as word stress and neighboring

sounds (i.e., coarticulation; see, e.g., Samuel & Frost, 2015) and they seemed to be more relevant prompts than isolated phonemes.

The translation of the target words was prepared using three dictionaries: *Longman English-Japanese Dictionary*, *Eijiro*, and *Luminous English-Japanese Dictionary*. The translation that appeared as the first entry (primary meaning) in multiple dictionaries was selected for the tests. After all the translation was prepared, a Japanese-English bilingual was consulted to make sure that the translation was appropriate.

Table 4

Sets of Nonwords

Set X	Set Y
dringe	brenge
scroost	scrount
brelf	brelt
dwirst	droist
zasked	wasked
konth	bonth
phrosque	phlusque
frealt	phrolt
strulge	strilge
phlinx	phrynx
brolve	swolve
heak	heem
scrisque	strasque
thwooth	threath
valph	twaph
threnge	grounge
straunch	splaunch
clend	drend
fleave	flarse
sploast	splaunt
Average number of phonemes in the set (*SD*)	
5.00	5.00
(0.86)	(0.86)
Average number of letters in the set (*SD*)	
6.25	6.25
(1.12)	(1.12)

2.2.5 Vocabulary Test Procedure

To average out the effect of the sets of target words, the sets were assigned to the tests across lists using Latin square design (see Table 5). The lists were randomly assigned to the participants.

Table 5

The Way in Which Sets of Target Words Were Assigned to Tests Across Lists

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
List 1	A	B	C	D	E	F	G	H
List 2	H	A	B	C	D	E	F	G
List 3	G	H	A	B	C	D	E	F
List 4	F	G	H	A	B	C	D	E
List 5	E	F	G	H	A	B	C	D
List 6	D	E	F	G	H	A	B	C
List 7	C	D	E	F	G	H	A	B
List 8	B	C	D	E	F	G	H	A

Note. The letters from A to H indicate unique sets of target words. Tests 1 to 8 indicate different types of vocabulary tests (e.g., written form recall, aural meaning selection).

The order of the vocabulary tests and the order of the test items within each of the tests were randomized across the participants. For each test item there was a time limit after which the next item appeared automatically: 15 seconds for production tests (form recall and meaning recall tests) and 10 seconds for multiple-choice tests (yes-no and meaning selection tests). This amount of time was considered enough for answering the questions based on a pilot study. Test instructions were provided before each test in the participants' L1 so that they could understand the test formats well. The test instructions were followed by a few practice items and their answer keys. All the tests were programmed in an online platform, Qualtrics.

2.3 Reading and Listening Comprehension Tests

The reading and listening comprehension tests were constructed using practice tests of the Test of English for International Communication (TOEIC), a standardized test of English language proficiency for non-native speakers developed by Educational Testing Service. This

test is arguably the most widely used test of English in Japan, with over two million people taking the test in the nation each year (IIBC, 2022). The test scores have been used in high-stakes situations such as school admissions and employment. Given this wide use in the country, the test seemed to be appropriate for measuring the reading and listening comprehension skills of the target population in this study.

This standardized test of English was selected for this study to have a reliable measure of comprehension skills. I am aware, however, that the test also has limitations in terms of generalizability in the sense that its listening test materials were scripted. As scripted speech is qualitatively different from naturally occurring speech (Biber et al., 2004; Redeker, 1984), further evidence is needed as to the degree to which listening skills assessed with scripted speech may be generalized to listening skills in the real world. Ideally, more spontaneous speech should be used for listening comprehension tests in future studies (see, e.g., Clark, 2014 for the use of semi-scripted speech for a listening comprehension test).

In order to include different types of comprehension skills in the reading and listening tests, TOEIC test items were examined and three types were identified: (1) referential (literal), (2) referential (paraphrase), and (3) inferential. The first type, referential (literal), was close to the propositional level of comprehension, that is, "a literal interpretation of the speaker's words" (Field, 2013, p. 100). The test items of this type could be answered by simply finding the relevant information (see the following subsections for example items of this type and the other two). Like the first type, the test items of the second type, referential (paraphrase), could be answered by understanding what was said explicitly, but the answer key of the comprehension question was paraphrased and so finding the exact word or phrase was not enough. For the third type, inferential, the literal understanding of words, phrases, or sentences was insufficient to

answer the question, and the accurate interpretation needed to be reached using other sources such as pragmatic and contextual knowledge (Field, 2013).

2.3.1 Reading Test

The reading test consisted of 16 referential (literal) items, 16 referential (paraphrase) items, and 16 inferential items (48 items in total). All were multiple-choice items with four options. The participants read one-way (e.g., articles) and two- (or more) way (e.g., instant messages) texts and answered two to four questions about each. See Appendix B for an example text and questions. This example text is an online chat discussion in a business context. The second item is referential (literal) because the answer (*On Tuesday*) to the question (*When will the crew begin work?*) can be found in the text (*… we could actually begin the job on Tuesday*). The third item is referential (paraphrase) because the answer (*Contact the client*) to the question (*What will Ms. Vega most likely do next?*) can be found in the text but is paraphrased (*I'll call the client this morning and let them know.*). The first item is inferential because the answer (*A hotel chain*) to the question (*What kind of business does the client most likely own?*) cannot be found in the text and needs to be inferred from the entire text.

2.3.2 Listening Test

The listening test consisted of 16 referential (literal) items, 16 referential (paraphrase) items, and 16 inferential items (48 items in total). All were multiple-choice items with four options. The participants listened to conversations and monologues and answered three questions about each. The audio was played only once. See Appendix C for an example audio script and questions. This example script is a telephone message in a business context. The first item is referential (literal) because the answer (*Marketing*) to the question (*What department does the*

speaker work in?) can be found in the telephone message (*... our marketing department*). The second item is referential (paraphrase) because the answer (*He bought a gift.*) to the question (*What does the speaker say he did yesterday?*) can be found in the telephone message but is paraphrased (*I went to the store yesterday to purchase her favorite chocolates for us to give to her…*). The third item is inferential because the answer (*The listener should give a speech.*) to the question (*What does the speaker imply when he says, "you've worked with her the most"?*) cannot be found in the telephone message and needs to be inferred from the entire message.

2.3.3 Reading and Listening Test Procedure

The reading and listening tests were timed following the time limits usually given in the TOEIC test. The reading test took about 60 minutes, and the listening test took about 25 minutes. The order of the two tests and the order of the blocks of items in each of the tests were randomized across the participants. Test instructions were provided before each test in the participants' L1 so that they could understand the test procedure well. The test instructions were followed by a practice block of items and their answer keys. As with the vocabulary tests, the tests were programmed in Qualtrics.

2.4 General Procedure

The participants completed two sessions, one for the vocabulary tests, and another for the reading and listening tests. Before the study, the participants completed a consent form and background questionnaire in their L1. See Appendix D for the translation of the background questionnaire. In the first session, the participants individually met with the researcher through video conferencing with Zoom and completed the eight vocabulary tests. As described earlier, one of the eight counter-balanced lists was randomly assigned to the participants for the

vocabulary tests. There was a 5-minute break after the fourth test. The first session took about 90 minutes. In the second session, the participants completed the reading and listening tests online by themselves. There was a 5-minute break between the two tests. The second session also took about 90 minutes. To reduce fatigue, the two sessions took place on separate days. The participants' responses to the form recall test of aural vocabulary were audio-recorded. The other (non-aural) data was recorded by Qualtrics.

2.5 Analysis

2.5.1 Scoring

All the test items were scored dichotomously as either correct (1) or incorrect (0) using the procedures described in the following subsections. The scoring was conducted by myself (a proficient speaker of English; L1 Japanese), although part of the data was scored by others as well to calculate interrater reliability (see the following subsections). Each test score was expressed as an accuracy percentage. The internal consistency of the tests was calculated by Cronbach's alpha. If some items were negatively correlated with the total scale and undermining the internal consistency of the test, those items were removed from the analysis. To ensure that vocabulary test items from the four frequency bands equally contributed to the test score (i.e., 25% each) in all tests in all lists even after item removal, a percent score was first calculated for each of the four frequency bands and then a total percent score was calculated by taking the average of the four percent scores.[1]

[1] Per a committee member's suggestion, I conducted a follow-up analysis where z-scores of different frequency bands were calculated and used for the analysis. The main findings were the same as those of the original analysis. See Appendix E for the results.

2.5.1.1 Form Recall Test Scoring

The participants' response was marked correct if the target word was produced. Even if the response included minor misspelling or mispronunciation, it was marked correct as long as it was recognized as the target word and it was not confused with another word. See Appendix F for more on the scoring guidelines. To calculate interrater reliability, a native English speaker also scored 25% of the data. This native rater was not familiar with a Japanese accent. Cohen's kappa was used for interrater reliability estimates.

2.5.1.2 Meaning Recall Test Scoring

The participants' response was marked correct if L1 meaning of the target word was provided. See Appendix G for more on the scoring guidelines. To calculate interrater reliability, another Japanese-English bilingual also scored 25% of the data. Cohen's kappa was used for interrater reliability estimates.

2.5.1.3 Yes-No Test Scoring

The hits-minus-false-alarms rule was used to score the yes-no tests (see, e.g., Hui & Godfroid, 2021 for a previous study with this scoring method). "Yes" responses to target words were coded as hits, and "Yes" responses to nonwords were coded as false alarms. The percentage of false alarms was subtracted from the percentage of hits. Although more complex scoring methods exist, for example, methods based on signal detection theory and correction for guessing formulas, there does not seem to be a large difference in outcomes between those complex methods and the simple hits-minus-false-alarms method (Mochida & Harrington, 2006).

2.5.1.4 Meaning Selection Test and Comprehension Test Scoring

The participants' response was marked correct if the correct answer option was selected.

2.5.2 Statistical Procedures

Before conducting the analyses, test scores were inspected for outliers. For each test, scores that fell outside the range of 2.5 standard deviations from the mean were considered outliers. These data points were removed from the analysis.

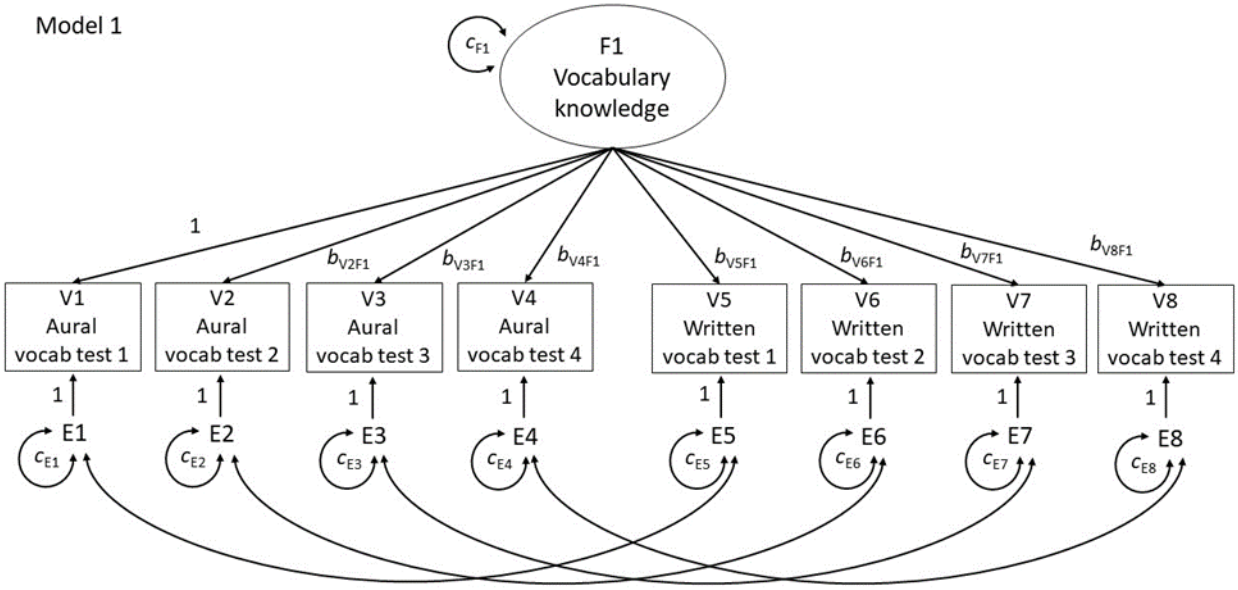
As a preliminary analysis, descriptive statistics were calculated for all the tests. Pearson's correlations among the variables were also computed.

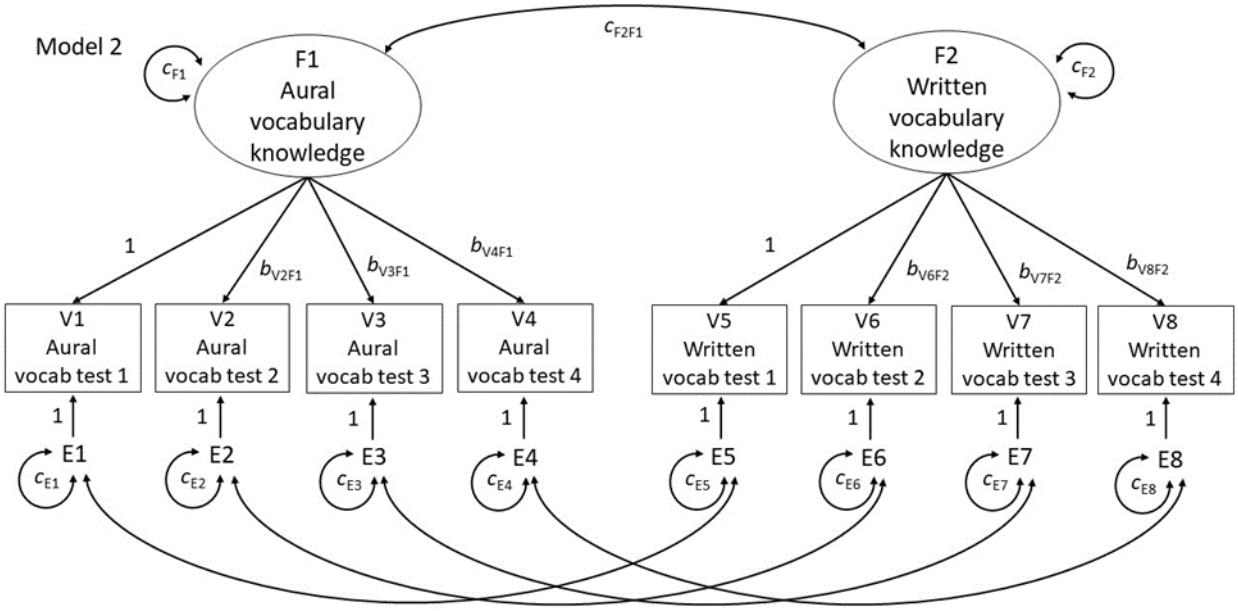
The main analysis was conducted with SEM using Mplus Version 8.11 (Muthén & Muthén, 2017). Unless otherwise stated, for each model I provided the following model fit indices: a model χ^2 statistic, Standardized Root Mean Squared Residual (SRMR), Root Mean Squared Error of Approximation (RMSEA), Akaike Information Criterion (AIC), and Comparative Fit Index (CFI). Following Hu and Bentler's (1999) guidelines, a model was considered acceptable when SRMR $\leq .08$, RMSEA $\leq .06$, and CFI $\geq .95$. Maximum likelihood estimation with Satorra-Bentler corrections was used to address the potential violation of the assumption of multivariate normality. Because of these adjustments Satorra-Bentler scaled chi-square difference testing was used for nested model comparison. Construct reliability was evaluated with Coefficient *H* (Hancock & Mueller, 2001). All formal statistical tests were based on unstandardized values, and standardized values were reported for interpretation purposes where appropriate.

To answer the first research question, the best vocabulary knowledge structure was identified by comparing three confirmatory factor models (Figure 2). In all the models, factors were scaled by fixing the first loading to 1, and four error covariances were included for parallel test formats. Model 1 ($df = 16$) was a one-factor model representing a construct of general

vocabulary knowledge (with eight indicators). Model 2 (*df* = 15) was a two-factor model representing constructs of aural and written vocabulary knowledge (each with four indicators). Aural and written vocabulary knowledge factors were allowed to covary because it was reasonable to assume that these vocabulary constructs would overlap to some degree. Model 3 (*df* = 8) was a residualized-factor model representing a construct of general vocabulary knowledge (with eight indicators) and constructs of aural and written modality-specific skills (each with four indicators). The latter two factors were orthogonal to the general vocabulary knowledge and to each other to capture unique modality effects. Model 1 was nested within Model 2 and Model 3 and so these comparisons (i.e., Model 1 vs. Model 2; Model 1 vs. Model 3) were planned to be made by chi-square difference tests.[2] If the difference was statistically significant, the more complex model would be considered a better model. If not significant, the less complex model would be considered better for its parsimony. Models 2 and 3 were not nested and so they were planned to be compared descriptively using model fit indices.

[2] Model 1 (the one-factor model) was nested within Model 2 (the oblique two-factor model) because the former was a special case of the latter where the two factors were perfectly correlated (Kline, 2016).


Model 1
F1
Vocabulary
knowledge
V1
Aural
vocab test 1
V2
Aural
vocab test 2
V3
Aural
vocab test 3
V4
Aural
vocab test 4
V5
Written
vocab test 1
V6
Written
vocab test 2
V7
Written
vocab test 3
V8
Written
vocab test 4
E1
E2
E3
E4
E5
E6
E7
E8


Model 2
F1
Aural
vocabulary
knowledge
F2
Written
vocabulary
knowledge
V1
Aural
vocab test 1
V2
Aural
vocab test 2
V3
Aural
vocab test 3
V4
Aural
vocab test 4
V5
Written
vocab test 1
V6
Written
vocab test 2
V7
Written
vocab test 3
V8
Written
vocab test 4
E1
E2
E3
E4
E5
E6
E7
E8

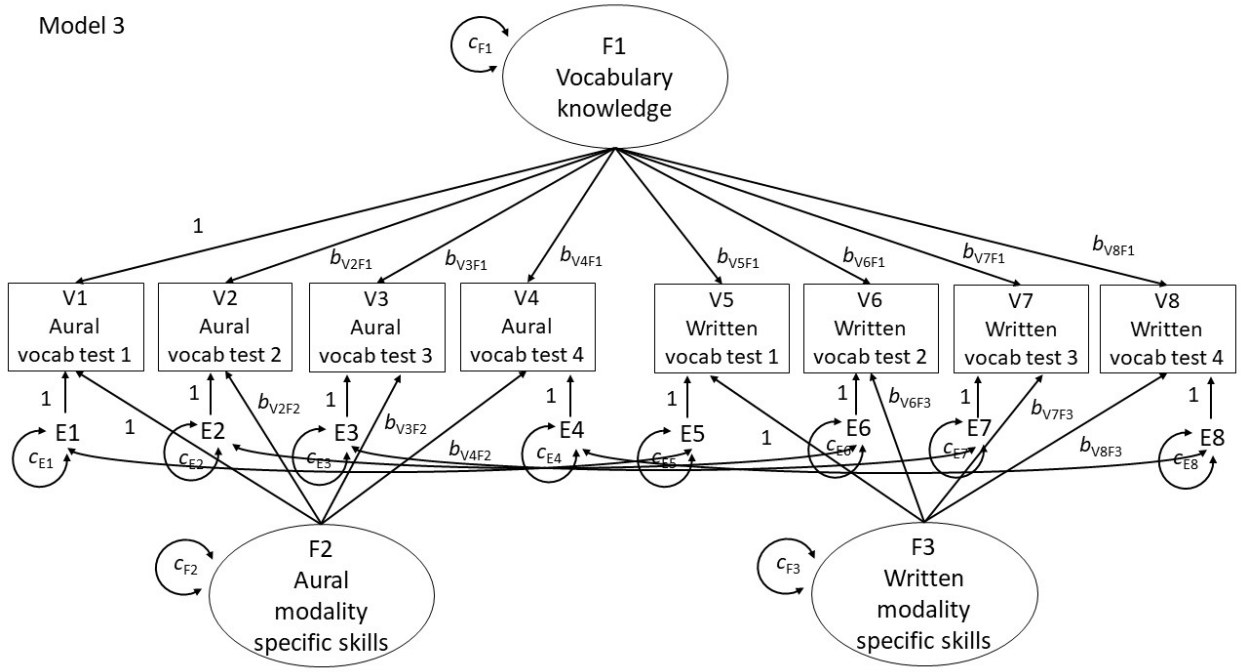


Figure 2. Three competing models of vocabulary knowledge structure.

The second research question concerned the contribution of modality-specific aspects to listening and reading comprehension. Since this analysis depended on the result of the preceding analysis on vocabulary structure, the analysis procedures were planned for each situation.

If Model 3 was the best model, latent variable path analysis would be conducted with the residualized-factor vocabulary model, where the factors of vocabulary knowledge and aural modality-specific skills predicted a construct of listening comprehension skills, and the factors of vocabulary knowledge and written modality-specific skills predicted a construct of reading comprehension skills (Figure 3; *df* = 57). As to the portion of listening and reading skills, factors were scaled by fixing the first loading to 1, and three error covariances were included for parallel item types. Disturbances of the listening and reading comprehension factors were allowed to

covary because it was reasonable to assume that these comprehension constructs would overlap to some degree. The paths of interest were b_{F4F2} and b_{F5F3}; since the exogenous factors were orthogonal to one another, the path b_{F4F2} indicated whether the factor of aural modality-specific skills explained the factor of listening comprehension skills above and beyond the vocabulary knowledge factor, and the path b_{F5F3} indicated whether the factor of written modality-specific skills explained the factor of reading comprehension skills above and beyond the vocabulary knowledge factor. It was planned to be examined whether these paths were statistically significant and how much they would explain (i.e., R^2).

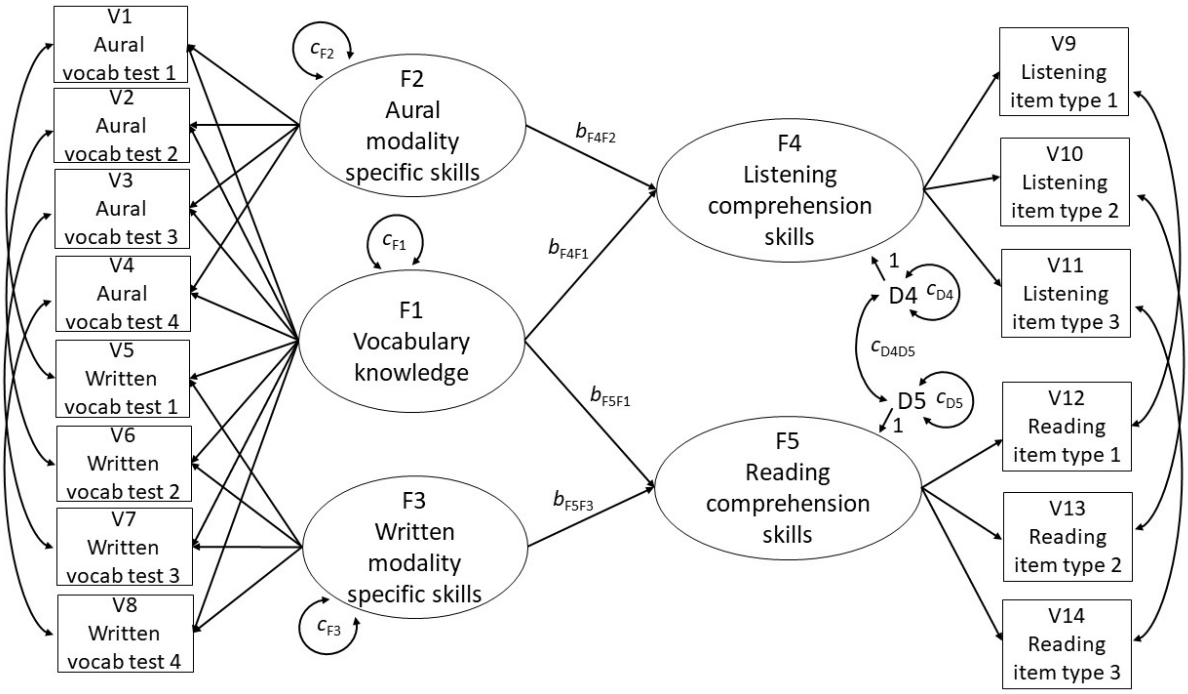


Figure 3. Latent variable path analysis model with the residualized-factor vocabulary knowledge structure explaining listening and reading comprehension skills.

If Model 2 was the best model, latent variable path analysis would be conducted with the two-factor vocabulary model (Figure 4; $df = 64$). The listening and reading portion was the same as in the previous model. In this model each of the listening and reading comprehension factors was explained by the two factors, aural and written vocabulary knowledge.[3] To estimate the unique contribution of aural vocabulary knowledge to listening comprehension, delta R^2 would be calculated, that is, the variance in listening comprehension skills explained by the two factors minus the variance explained by the written vocabulary factor alone. Similarly, to estimate the unique contribution of written vocabulary knowledge to reading comprehension, delta R^2 would be calculated, that is, the variance in reading comprehension skills explained by the two factors minus the variance explained by the aural vocabulary factor alone.

[3] There are cross paths (i.e., from aural vocabulary to reading and from written vocabulary to listening) in this model because the two vocabulary knowledge factors are correlated and both include vocabulary knowledge in general in addition to modality-specific knowledge. Without those paths, the data-model fit would be bad because no path means no relation between variables.

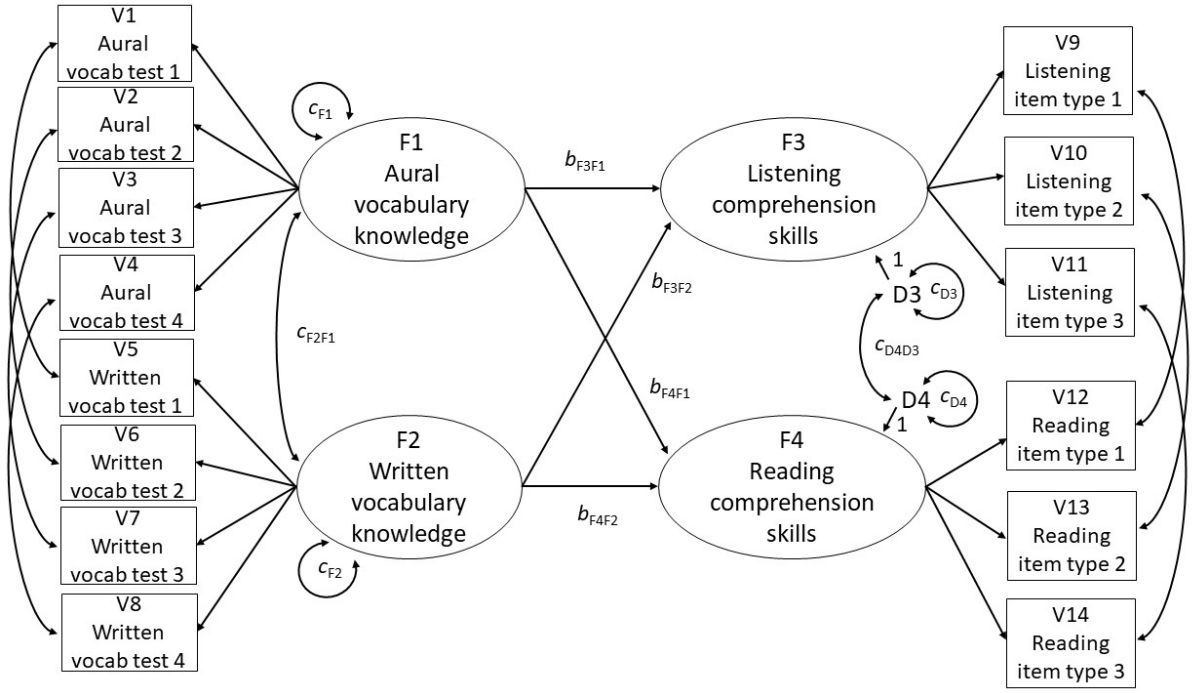


Figure 4. Latent variable path analysis model with the two-factor vocabulary knowledge structure explaining listening and reading comprehension skills.

If Model 1 was the best model, latent variable path analysis would be conducted with the one-factor vocabulary model (Figure 5; *df* = 67). The listening and reading portion was the same as in the previous models. It was planned to be examined how much the factor of vocabulary knowledge would explain the listening and reading comprehension factors. Even though this should and would be the primary analysis in the case that Model 1 was the best model, I would also look at the second-best model (one of the other two scenarios described in the preceding paragraphs) for exploratory purposes.

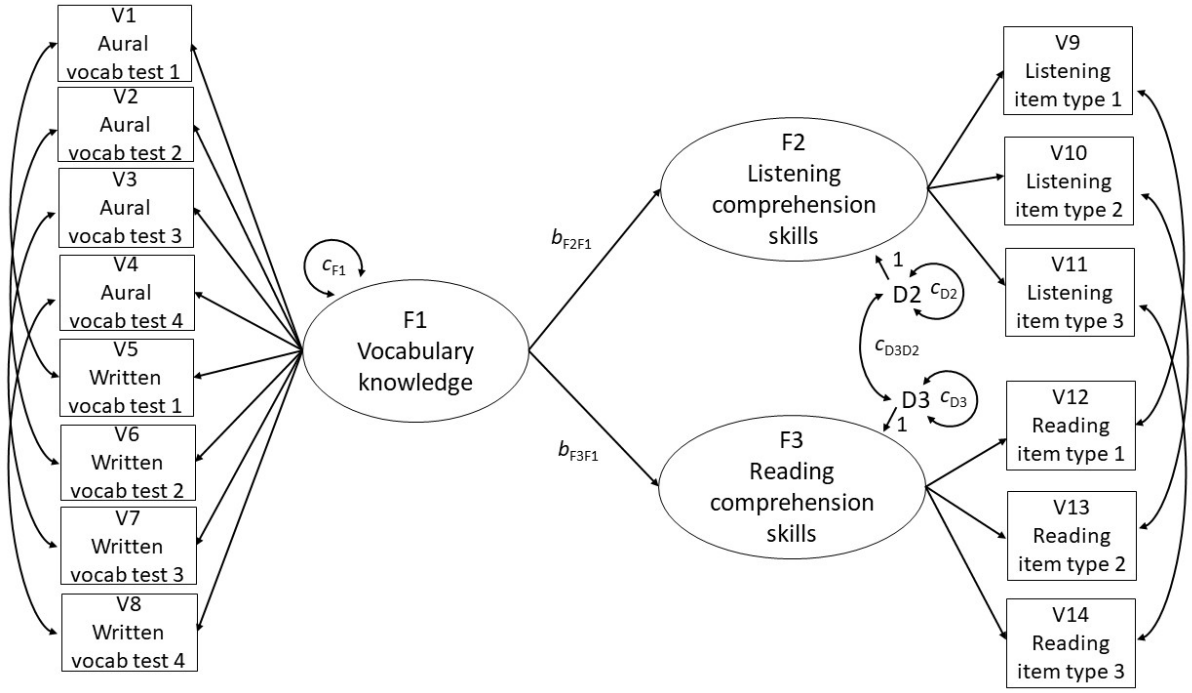


Figure 5. Latent variable path analysis model with the one-factor vocabulary knowledge structure explaining listening and reading comprehension skills.

The third research question concerned the relative levels of aural and written vocabulary knowledge. This analysis would be conducted only if the best model of vocabulary knowledge structure consisted of aural and written vocabulary (i.e., Model 2). In this analysis latent means modeling would be used to examine whether the participants' aural vocabulary knowledge was larger, smaller, or equivalent to their written vocabulary knowledge. The model was basically the same as Model 2, but the mean structure was modeled in addition to the covariance structure (Figure 6; $df = 21$). The parameters a_{F1} and a_{F2} represented factor means for aural and written vocabulary knowledge, respectively. The parameter a_{F1} was set to 0 as a reference, and so the parameter a_{F2} would indicate the factor mean of written vocabulary knowledge relative to aural

vocabulary (i.e., difference between aural and written vocabulary knowledge). Given the parallel test formats, loadings and intercepts were assumed to be invariant. These constraints would be examined stepwise, first for loadings and then for intercepts. If loading(s) or intercept(s) were judged to be non-invariant, the constraints would be released. Even under a partial non-invariant condition, factor means are still comparable (G. Hancock, personal communication, March 8, 2023; see also Hancock et al., 2009, where it is argued that latent mean inference remains uncompromised even under fairly minimal invariance conditions).[4] Thus, if invariance or partial invariance held, I would compare the factor means. (If not, I would only look at means at the level of measured variables.) It was planned to be examined whether the parameter α_{F2} was statistically significant and which direction (positive or negative). The magnitude of the factor mean difference would be evaluated with estimated standardized effect size (Hancock, 2001). Also, when a model includes a mean structure, incremental fit indices such as CFI are ill-advised because of the unclear baseline model (Thompson & Green, 2013; Widaman & Thompson, 2003), and therefore they would not be used for the assessment of the latent means model.

[4] In the world of measured variables, scores are the aggregations of equally-contributing items. In the world of latent constructs, on the other hand, scores are the representations of invisible factors manifesting into (or indicated by) observed variables. Here, the fact that some observed variables are better indicators of one factor than of the other (i.e., partial noninvariance) does not entirely diminish the comparability of the factors (G. Hancock, personal communication, July 10, 2023).

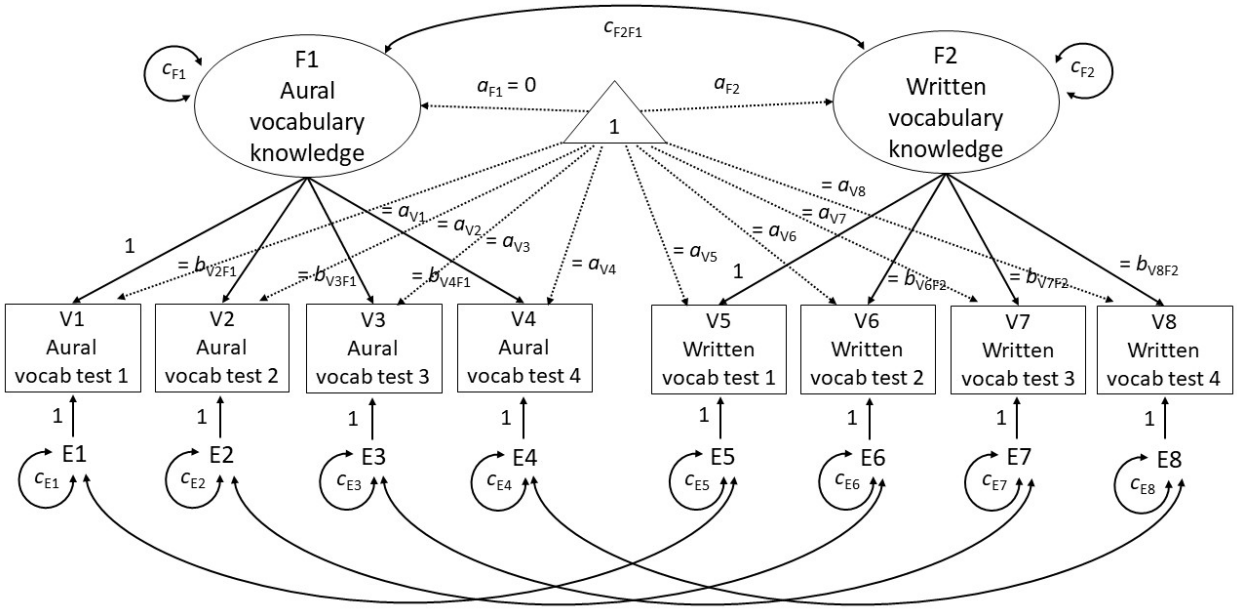


Figure 6. Latent means model of aural and written vocabulary knowledge.

Chapter 3: Results

3.1 Preliminary Analysis

3.1.1 Form Recall Test of Written Vocabulary

First, I scored the participants' responses by myself. After that, another rater, who was a native speaker of English, scored 25% of the responses. The interrater reliability measured by Cohen's kappa was .94, which was in the range of "almost perfect" agreement on the basis of Landis and Koch's (1977) guidelines, suggesting that the scoring was reasonable.

Next, test item statistics were examined to check for malfunctioning items. The following items were negatively correlated with the total scale: *ocean*, *secret*, *employ*, *wolf* in List 1, *pencil* in List 2, *intensive* in List 3, *exit*, *vinegar*, *destiny* in List 5, *onion* in List 6, *honest*, *cellular* in List 7, *hobby*, *poem* in List 8. These items (about 4% of all items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of the test were .91, .90, .91, .93, .85, .91, .92, and .92, respectively.

Then, a percentage accuracy score was calculated for each participant. The average score was 52.29 (SD = 21.54). None of the scores fell outside the range of 2.5 standard deviations from the mean. The descriptive statistics for this test were: N = 185, M = 52.29, SD = 21.54, range = (0, 97.50), skew = -0.21, kurtosis = -0.77.

3.1.2 Form Recall Test of Aural Vocabulary

One participant's response was lost due to a technical problem. This data point (about 0.5% of the data) was treated as a missing value.

First, I scored the participants' responses by myself. After that, another rater, who was a native speaker of English, scored 25% of the responses. The interrater reliability measured by Cohen's kappa was .81, which was in the range of "almost perfect" agreement on the basis of Landis and Koch's (1977) guidelines, suggesting that the scoring was reasonable.

Next, test item statistics were examined to check for malfunctioning items. The following items were negatively correlated with the total scale: *solar*, *pepper*, *fossil* in List 2, *tense* in List 3, *temple*, *tragic*, *collide* in List 5, *treasure*, *ultimate* in List 6, *decay* in List 7, *salt*, *embed* in List 8. These items (about 4% of all items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of the test were .84, .92, .90, .92, .87, .89, .91, and .88, respectively.

Then, a percentage accuracy score was calculated for each participant. The average score was 43.07 (SD = 20.09). None of the scores fell outside the range of 2.5 standard deviations from the mean. The descriptive statistics for this test were: N = 184, M = 43.07, SD = 20.09, range = (2.78, 87.50), skew = -0.01, kurtosis = -0.74.

3.1.3 Meaning Recall Test of Written Vocabulary

First, I scored the participants' responses by myself. After that, another rater, who was a Japanese-English bilingual, scored 25% of the responses. The interrater reliability measured by Cohen's kappa was .95, which was in the range of "almost perfect" agreement on the basis of Landis and Koch's (1977) guidelines, suggesting that the scoring was reasonable.

Next, test item statistics were examined to check for malfunctioning items. The following items were negatively correlated with the total scale: *prison*, *manuscript*, *cellular* in List 1, *hobby* in List 2, *award*, *harbor* in List 5, *diary*, *legal* in List 6, *vinegar* in List 7, *onion* in List 8.

These items (about 3% of all items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of the test were .92, .94, .94, .94, .92, .89, .92, and .94, respectively.

Then, a percentage accuracy score was calculated for each participant. The average score was 61.50 (*SD* = 22.56). One participant's score fell outside the range of 2.5 standard deviations from the mean. This data point (about 0.5% of the data) was removed from the analysis as an outlier. The descriptive statistics for this test were: *N* = 184, *M* = 61.83, *SD* = 22.20, range = (10.00, 100), skew = -0.40, kurtosis = -0.75.

3.1.4 Meaning Recall Test of Aural Vocabulary

First, I scored the participants' responses by myself. After that, another rater, who was a Japanese-English bilingual, scored 25% of the responses. The interrater reliability measured by Cohen's kappa was .96, which was in the range of "almost perfect" agreement on the basis of Landis and Koch's (1977) guidelines, suggesting that the scoring was reasonable.

Next, test item statistics were examined to check for malfunctioning items. The following items were negatively correlated with the total scale: *justice* in List 1, *length* in List 2, *secret* in List 4, *escape*, *poison*, *ladder* in List 5, *effect*, *educate* in List 6, *diary*, *homework* in List 7. These items (about 3% of all items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of the test were .88, .92, .94, .94, .87, .91, .90, and .90, respectively.

Then, a percentage accuracy score was calculated for each participant. The average score was 44.38 (*SD* = 21.34). None of the scores fell outside the range of 2.5 standard deviations from

the mean. The descriptive statistics for this test were: $N = 185$, $M = 44.38$, $SD = 21.34$, range = (0, 92.22), skew = -0.01, kurtosis = -0.82.

3.1.5 Yes-No Test of Written Vocabulary

First, test item statistics were examined to check for malfunctioning items. The following real word items were negatively correlated with the total scale: *rent*, *object*, *treasure*, *servant*, *vinegar* in List 1, *factor* in List 2, *monkey*, *shelf*, *windy* in List 4, *wolf* in List 5, *honey* in List 6, *award*, *theft* in List 7, *harvest*, *valley*, *annual* in List 8. These items (5% of all real word items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of real words of the test were .89, .94, .93, .94, .93, .88, .92, and .91, respectively. Similarly, test item statistics were examined for nonword items. None of the nonword items were negatively correlated with the total scale. The Cronbach's alpha estimate of internal consistency for the nonword items was .60.

Then, a percentage accuracy score was calculated for each participant. The average score of real word items was 78.18 ($SD = 18.88$). After subtracting the percentage of false alarms ("Yes" responses to nonwords), which was on average 2.35% ($SD = 4.97$), the average score was 75.82 ($SD = 19.84$). Three participants' scores fell outside the range of 2.5 standard deviations from the mean. These data points (about 2% of the data) were removed from the analysis as outliers. The descriptive statistics for this test were: $N = 182$, $M = 76.87$, $SD = 18.24$, range = (26.39, 100), skew = -0.91, kurtosis = 0.06.

3.1.6 Yes-No Test of Aural Vocabulary

First, test item statistics were examined to check for malfunctioning items. The following real word items were negatively correlated with the total scale: *gentle*, *grammar* in List 1,

treasure, *muddy* in List 2, *naked* in List 3, *delicious*, *divide*, *awake*, *arrow*, *odor*, *cellular* in List 4, *hobby*, *tiny*, *shelf*, *riot* in List 5, *interest* in List 6, *absorb*, *noisy*, *melt*, *dialect*, *voter* in List 7, *emotion*, *cruel* in List 8. These items (about 7% of all real word items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of real words of the test were .84, .88, .92, .85, .87, .88, .89, and .87, respectively. Similarly, test item statistics were examined for nonword items. None of the nonword items were negatively correlated with the total scale. The Cronbach's alpha estimate of internal consistency for the nonword items was .64.

Then, a percentage accuracy score was calculated for each participant. The average score of real word items was 69.26 (*SD* = 18.33). After subtracting the percentage of false alarms ("Yes" responses to nonwords), which was on average 14.57% (*SD* = 12.20), the average score was 54.69 (*SD* = 22.30). Two participants' scores fell outside the range of 2.5 standard deviations from the mean. These data points (about 1% of the data) were removed from the analysis as outliers. The descriptive statistics for this test were: *N* = 183, *M* = 55.38, *SD* = 21.42, range = (5.56, 100), skew = -0.23, kurtosis = -0.70.

3.1.7 Meaning Selection Test of Written Vocabulary

First, test item statistics were examined to check for malfunctioning items. The following items were negatively correlated with the total scale: *donate* in List 5, *poem*, *revive* in List 6, *needle* in List 7, *eliminate*, *voyage* in List 8. These items (about 2% of all items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of the test were .82, .93, .89, .90, .84, .88, .88, and .91, respectively.

Then, a percentage accuracy score was calculated for each participant. The average score was 87.72 (*SD* = 13.06). Six participants' scores fell outside the range of 2.5 standard deviations from the mean. These data points (about 3% of the data) were removed from the analysis as outliers. The descriptive statistics for this test were: *N* = 179, *M* = 88.99, *SD* = 11.18, range = (56.39, 100), skew = -0.97, kurtosis = -0.12.

3.1.8 Meaning Selection Test of Aural Vocabulary

First, test item statistics were examined to check for malfunctioning items. The following items were negatively correlated with the total scale: *cousin*, *narrow*, *pupil* in List 1, *harbor*, *saint* in List 2, *legal* in List 3, *historical*, *exit* in List 4, *exam*, *critical* in List 5, *hobby* in List 7, *absent*, *wolf* in List 8. These items (about 4% of all items) were removed from the analysis. After the removal, the Cronbach's alpha estimates of internal consistency for the eight lists of the test were .88, .90, .92, .91, .88, .82, .92, and .88, respectively.

Then, a percentage accuracy score was calculated for each participant. The average score was 81.34 (*SD* = 15.96). Four participants' scores fell outside the range of 2.5 standard deviations from the mean. These data points (about 2% of the data) were removed from the analysis as outliers. The descriptive statistics for this test were: *N* = 181, *M* = 82.36, *SD* = 14.56, range = (42.50, 100), skew = -1.00, kurtosis = 0.14.

3.1.9 Reading Comprehension Test

Five participants did not complete the reading comprehension test. These data points (about 3% of the data) were treated as missing values.

First, test item statistics were examined to check for malfunctioning items. One item in the inferential item type was negatively correlated with the total scale within the given item type.

This item (about 2% of all items) was removed from the analysis. After the removal, the internal consistency of the test as a whole measured by Cronbach's alpha was .86. The Cronbach's alpha estimates of internal consistency for the three types of items, referential (literal), referential (paraphrase), and inferential, were .70, .63, and .67, respectively.

Then, percentage accuracy scores were calculated for each participant. The average overall score was 66.93 (*SD* = 16.30). The average scores for the three types of items were 72.40 (*SD* = 18.20) for referential (literal), 61.42 (*SD* = 18.28) for referential (paraphrase), and 66.96 (*SD* = 18.27) for inferential. Two participants' scores fell outside the range of 2.5 standard deviations from the mean for the referential (literal) item type. These data points (about 1% of the data) were removed from the analysis as outliers. Similarly, one participant's score (about 0.5% of the data) in the referential (paraphrase) item type and three participants' scores (about 2% of the data) in the inferential item type were identified as outliers and were removed from the analysis. The descriptive statistics for the reading test were: *N* = 180, *M* = 67.13, *SD* = 15.89, range = (25.00, 100), skew = -0.41, kurtosis = -0.49 for the overall scores, *N* = 178, *M* = 72.96, *SD* = 17.48, range = (31.25, 100), skew = -0.48, kurtosis = -0.51 for the referential (literal) item type, *N* = 179, *M* = 61.70, *SD* = 17.96, range = (18.75, 100), skew = -0.16, kurtosis = -0.75 for the referential (paraphrase) item type, and *N* = 177, *M* = 67.80, *SD* = 17.24, range = (26.67, 100), skew = -0.39, kurtosis = -0.39 for the inferential item type.

3.1.10 Listening Comprehension Test

Five participants did not complete the listening comprehension test. These data points (about 3% of the data) were treated as missing values.

First, test item statistics were examined to check for malfunctioning items. One item in the referential (literal) item type and one item in the inferential item type were negatively correlated with the total scale within the given item types. These items (about 4% of all items) were removed from the analysis. After the removal, the internal consistency of the test as a whole measured by Cronbach's alpha was .85. The Cronbach's alpha estimates of internal consistency for the three types of items, referential (literal), referential (paraphrase), and inferential, were .61, .68, and .66, respectively.

Then, percentage accuracy scores were calculated for each participant. The average overall score was 60.21 (*SD* = 16.71). The average scores for the three types of items were 65.93 (*SD* = 17.94) for referential (literal), 56.60 (*SD* = 19.64) for referential (paraphrase), and 58.11 (*SD* = 19.92) for inferential. One participant's score fell outside the range of 2.5 standard deviations from the mean for the referential (literal) item type. This data point (about 0.5% of the data) was removed from the analysis as an outlier. Similarly, one participant's score (about 0.5% of the data) was identified as an outlier for each of the referential (paraphrase) and inferential item types and was removed from the analysis. The descriptive statistics for the listening test were: *N* = 180, *M* = 60.35, *SD* = 16.51, range = (21.81, 97.78), skew = 0.01, kurtosis = -0.73 for the overall scores, *N* = 179, *M* = 66.18, *SD* = 17.65, range = (26.67, 100), skew = -0.19, kurtosis = -0.73 for the referential (literal) item type, *N* = 179, *M* = 56.88, *SD* = 19.33, range = (12.50, 100), skew = -0.02, kurtosis = -0.65 for the referential (paraphrase) item type, and *N* = 179, *M* = 58.40, *SD* = 19.60, range = (13.33, 100), skew = -0.12, kurtosis = -0.61 for the inferential item type.

3.1.11 Summary of Preliminary Analysis

A summary of the descriptive statistics for the measured variables is shown in Table 6.[5] Pearson's correlations among the measured variables are shown in Table 7.

[5] There were a few scores at or below chance level. If I remove these scores and treat them as missing values, missingness is dependent on the outcome variables, which potentially distorts the data by removing information of low scores. Also, such low scores were only about 1% of all data points and were unlikely to affect the overall findings. That being said, a follow-up analysis was conducted without scores at or below chance level. The main findings were the same as those of the original analysis, which will be presented in the next section.

Table 6

Descriptive Statistics for the Measured Variables

Measure	*N*	Mean	*SD*	Min.	Max.	Skew	Kurtosis
Written Form	185	52.29	21.54	0.00	97.50	-0.21	-0.77
Aural Form	184	43.07	20.09	2.78	87.50	-0.01	-0.74
Written Meaning	184	61.83	22.20	10.00	100.00	-0.40	-0.75
Aural Meaning	185	44.38	21.34	0.00	92.22	-0.01	-0.82
Written Yes-No	182	76.87	18.24	26.39	100.00	-0.91	0.06
Aural Yes-No	183	55.38	21.42	5.56	100.00	-0.23	-0.70
Written Selection	179	88.99	11.18	56.39	100.00	-0.97	-0.12
Aural Selection	181	82.36	14.56	42.50	100.00	-1.00	0.14
Reading (RL)	178	72.96	17.48	31.25	100.00	-0.48	-0.51
Reading (RP)	179	61.70	17.96	18.75	100.00	-0.16	-0.75
Reading (I)	177	67.80	17.24	26.67	100.00	-0.39	-0.39
Listening (RL)	179	66.18	17.65	26.67	100.00	-0.19	-0.73
Listening (RP)	179	56.88	19.33	12.50	100.00	-0.02	-0.65
Listening (I)	179	58.40	19.60	13.33	100.00	-0.12	-0.61

Note. Written Form = form recall test of written vocabulary (written vocab test 1); Aural Form = form recall test of aural vocabulary (aural vocab test 1); Written Meaning = meaning recall test of written vocabulary (written vocab test 2); Aural Meaning = meaning recall test of aural vocabulary (aural vocab test 2); Written Yes-No = yes-no test of written vocabulary (written vocab test 3); Aural Yes-No = yes-no test of aural vocabulary (aural vocab test 3); Written Selection = meaning selection test of written vocabulary (written vocab test 4); Aural Selection = meaning selection test of aural vocabulary (aural vocab test 4); Reading (RL) = reading test referential (literal) items (reading item type 1); Reading (RP) = reading test referential (paraphrase) items (reading item type 2); Reading (I) = reading test inferential items (reading item type 3); Listening (RL) = listening test referential (literal) items (listening item type 1); Listening (RP) = listening test referential (paraphrase) items (listening item type 2); Listening (I) = listening test inferential items (listening item type 3).

Table 7

Correlation Matrix for the Measured Variables

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Written Form	-													
2. Aural Form	.85	-												
3. Written Meaning	.87	.84	-											
4. Aural Meaning	.86	.86	.87	-										
5. Written Yes-No	.73	.70	.81	.76	-									
6. Aural Yes-No	.73	.73	.79	.81	.73	-								
7. Written Selection	.74	.73	.82	.77	.72	.65	-							
8. Aural Selection	.77	.77	.83	.79	.76	.73	.81	-						
9. Reading (RL)	.60	.59	.58	.53	.53	.49	.47	.57	-					
10. Reading (RP)	.64	.63	.62	.58	.49	.52	.52	.59	.70	-				
11. Reading (I)	.55	.53	.55	.53	.47	.48	.46	.49	.67	.65	-			
12. Listening (RL)	.46	.46	.42	.43	.38	.41	.31	.43	.57	.57	.48	-		
13. Listening (RP)	.61	.61	.56	.58	.47	.56	.40	.56	.65	.60	.59	.64	-	
14. Listening (I)	.70	.64	.62	.66	.52	.55	.54	.61	.62	.60	.57	.56	.69	-

Note. Correlations computed with pairwise deletion. All values are significant at $p < .01$. Written Form = form recall test of written vocabulary (written vocab test 1); Aural Form = form recall test of aural vocabulary (aural vocab test 1); Written Meaning = meaning recall test of written vocabulary (written vocab test 2); Aural Meaning = meaning recall test of aural vocabulary (aural vocab test 2); Written Yes-No = yes-no test of written vocabulary (written vocab test 3); Aural Yes-No = yes-no test of aural vocabulary (aural vocab test 3); Written Selection = meaning selection test of written vocabulary (written vocab test 4); Aural Selection = meaning selection test of aural vocabulary (aural vocab test 4); Reading (RL) = reading test referential (literal) items (reading item type 1); Reading (RP) = reading test referential (paraphrase) items (reading item type 2); Reading (I) = reading test inferential items (reading

item type 3); Listening (RL) = listening test referential (literal) items (listening item type 1); Listening (RP) = listening test referential (paraphrase) items (listening item type 2); Listening (I) = listening test inferential items (listening item type 3).

3.2 Main Analysis

3.2.1 RQ1: The Structure of Vocabulary Knowledge

The first research question asked what type of vocabulary knowledge structure would best fit the data. To this end, three confirmatory factor models were examined.

Model 1, a one-factor structure with general vocabulary, had the following results of data-model fit: $\chi^2 = 36.62$ ($df = 16$, $p = .002$), SRMR = .017, RMSEA = .083 with 90% CI [.048, .119], AIC = 10995.58, and CFI = .987. Using Hu and Bentler's (1999) guidelines (i.e., acceptable when SRMR ≤ .08, RMSEA ≤ .06, and CFI ≥ .95), the SRMR and CFI values were acceptable, whereas the RMSEA value was not in the recommended range.

Model 2, a two-factor structure with aural and written vocabulary, had the following results of data-model fit: $\chi^2 = 29.36$ ($df = 15$, $p = .014$), SRMR = .017, RMSEA = .072 with 90% CI [.031, .110], AIC = 10989.51, and CFI = .991. As with Model 1, the SRMR and CFI values were acceptable. The RMSEA value was better than that of Model 1, but it was slightly above the recommended range.

Model 3, a residualized-factor structure with general vocabulary and aural and written modality-specific skills, had an issue of non-convergence. By changing the factor scaling method (i.e., fixing factor variance or another loading to 1), this problem was solved. However, even with these alternative scaling methods, there was another problem in this model; the residual covariance matrix was not positive definite with a negative residual variance. This problem occurs often due to some model misspecification (Geiser, 2012; Muthén, 2012). When the model was examined to explore the culprit, it was found that none of the indicators loaded onto the residualized factors (i.e., modality-specific skill factors). This appeared to suggest over-

parameterization with too many factors in the model, leading to an improper solution. Therefore, Model 3 did not seem to be appropriate.

To compare Model 1 and Model 2, a Satorra-Bentler scaled chi-square difference test was conducted. The difference was statistically significant, $\Delta\chi^2_{(SB)} = 7.98$ ($df = 1$, $p = .005$), suggesting that Model 2 had a significantly better fit to the data than Model 1. Model 2 looked better than Model 1 based on other model fit indices also (see Table 8 for a summary of the model comparison).

Table 8

Summary of Model Comparison

Model	χ^2	*df*	$\Delta\chi^{2a}$	SRMR	RMSEA	AIC	CFI
Model 1 (one-factor structure)	36.62	16	-	.017	.083	10996	.987
Model 2 (two-factor structure)	29.36	15	7.98*	.017	.072	10990	.991

Note. Model 3 (residualized-factor structure) did not work and resulted in an improper solution.

*$p < .05$.

[a] The Satorra-Bentler scaled chi-square difference test.

Figure 7 shows the standardized parameter estimates for Model 2 (two-factor structure). Both factors had strong loadings from their indicators with the loadings 0.84 or higher.[6] The construct reliability measured by Coefficient *H* (Hancock & Mueller, 2001) was .95 for the aural

[6] There was a counterintuitive negative residual correlation for the meaning recall tests. This might be due to the very small residual variances of the meaning recall test scores. Because there was very little left unexplained in the meaning recall test scores, the residual correlation might not be open to meaningful interpretation.

vocabulary knowledge factor, and .96 for the written vocabulary knowledge factor. Importantly, however, the two factors were highly correlated (r = 0.98), indicating that the two factors overlapped to a considerable degree.

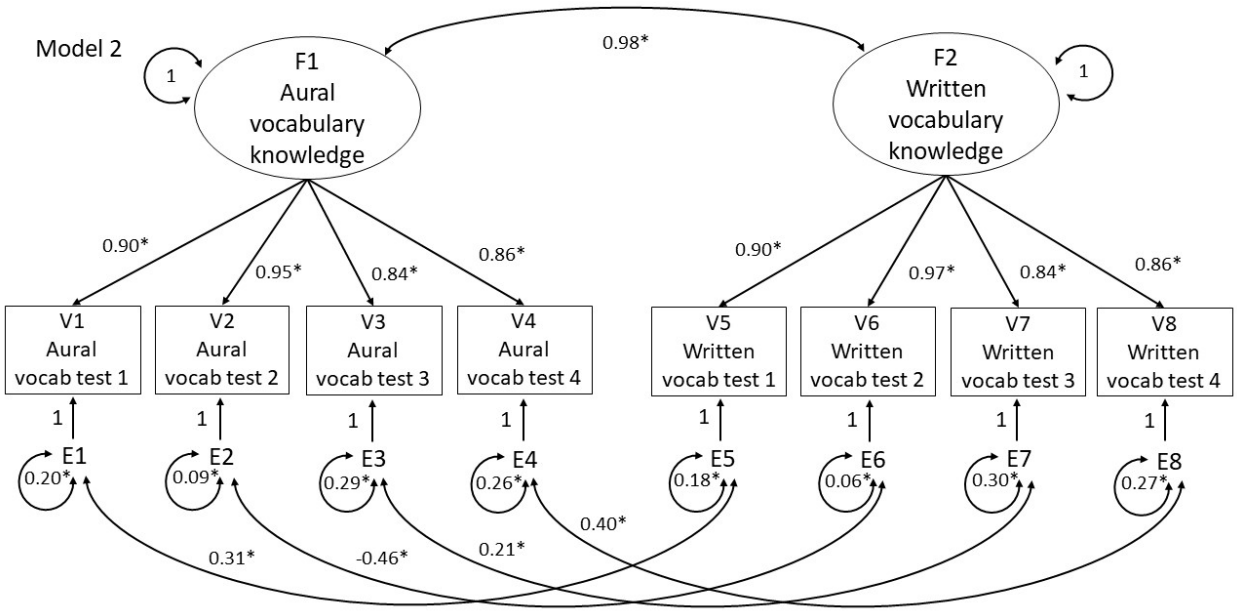


Figure 7. Standardized parameter estimates for the two-factor model of vocabulary knowledge structure. *p < .05.

Figure 8 shows the standardized parameter estimates for Model 1 (one-factor structure). The factor had strong loadings from the indicators with the loadings 0.83 or higher. The construct reliability was .98.[7]

[7] There is more than one way to model the data structure. In the analysis reported in this section, factor loadings were freely estimated for each construct, which have the advantage of getting closer to the reality by having less constraints. However, since parallel test formats were used across constructs, it is also possible to try to impose loading constraints across constructs. The

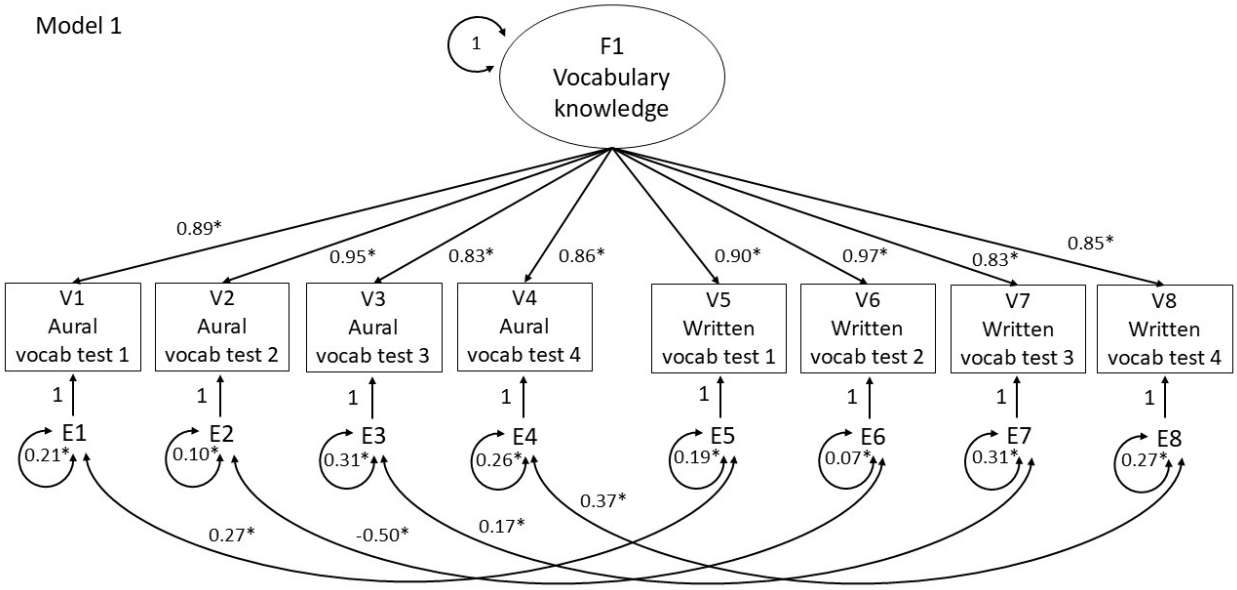


Figure 8. Standardized parameter estimates for the one-factor model of vocabulary knowledge structure. *$p < .05$.

3.2.2 RQ2: The Unique Contribution of Modality to Comprehension

The second research question asked the extent to which the modality-specific aspects of vocabulary knowledge explained listening and reading comprehension skills. Because in the preceding analysis a chi-square difference test showed a better data-model fit for the two-factor model over the one-factor model, but there was high factor correlation in the two-factor model, latent variable path analysis was first conducted with the two-factor vocabulary knowledge structure and then with the one-factor vocabulary knowledge structure as well.

results with this constraining approach are found in Appendix H. The main findings were the same as those of the original analysis.

Using Hu and Bentler's (1999) guidelines (i.e., acceptable when SRMR ≤ .08, RMSEA ≤ .06, and CFI ≥ .95), the data-model fit of the latent variable path model with the two-factor vocabulary knowledge structure was acceptable, $\chi^2 = 102.51$ ($df = 64$, $p = .002$), SRMR = .033, RMSEA = .057 with 90% CI [.035, .077], AIC = 19518.17, and CFI = .984. All the factors had strong loadings from their indicators with the standardized loadings 0.72 or higher. The construct reliability measured by Coefficient *H* (Hancock & Mueller, 2001) was .95 for the aural vocabulary knowledge factor, .96 for the written vocabulary knowledge factor, .85 for the listening comprehension skill factor, and .87 for the reading comprehension skill factor.

Figure 9 shows the standardized parameter estimates for the structural portion of the model. The structural path from the aural vocabulary knowledge factor to the listening comprehension skill factor was positive and statistically significant. The rest of the structural paths were not significant. It may be useful to note that the significant path coefficient was greater than 1, which may be somewhat rare for a standardized coefficient. This was probably due to a suppression effect caused by high correlation between the two vocabulary factors and was not an improper solution (see Geiser, 2022 in particular as well as Deegan, 1978). That said, to understand the influence of the high correlation between the predictors, a follow-up analysis was also conducted, where each of the predictors was included in the model as a single predictor. The follow-up analysis will be reported later in this section.

To estimate the unique contribution of aural vocabulary knowledge to listening comprehension above and beyond what written vocabulary knowledge could explain, delta R^2 was calculated, that is, the variance in listening comprehension skills explained by the two factors minus the variance explained by the written vocabulary factor alone. The delta R^2 was calculated by hand using the path tracing rules (see Appendix I for the derivation). The delta R^2

was 0.09, meaning that the aural vocabulary knowledge factor uniquely explained 9% of the variance in the listening comprehension skill factor. Using the same formula, the variance in listening uniquely explained by the written vocabulary factor was 2%, and since the variance explained by both factors was 62%, the common part of the two factors, which may be considered general vocabulary knowledge, explained 62 – (9 + 2) = 51% of the variance in the listening comprehension skill factor.

Similarly, the unique contribution of written vocabulary knowledge to reading comprehension was estimated using the same method as for listening. The variance in reading uniquely explained by the written vocabulary factor was only 1%. The variance in reading uniquely explained by the aural vocabulary factor was 0%. Since the variance explained by both factors was 56%, the common part of the two factors, which may be considered general vocabulary knowledge, explained 56 – (1 + 0) = 55% of the variance in the reading comprehension skill factor.

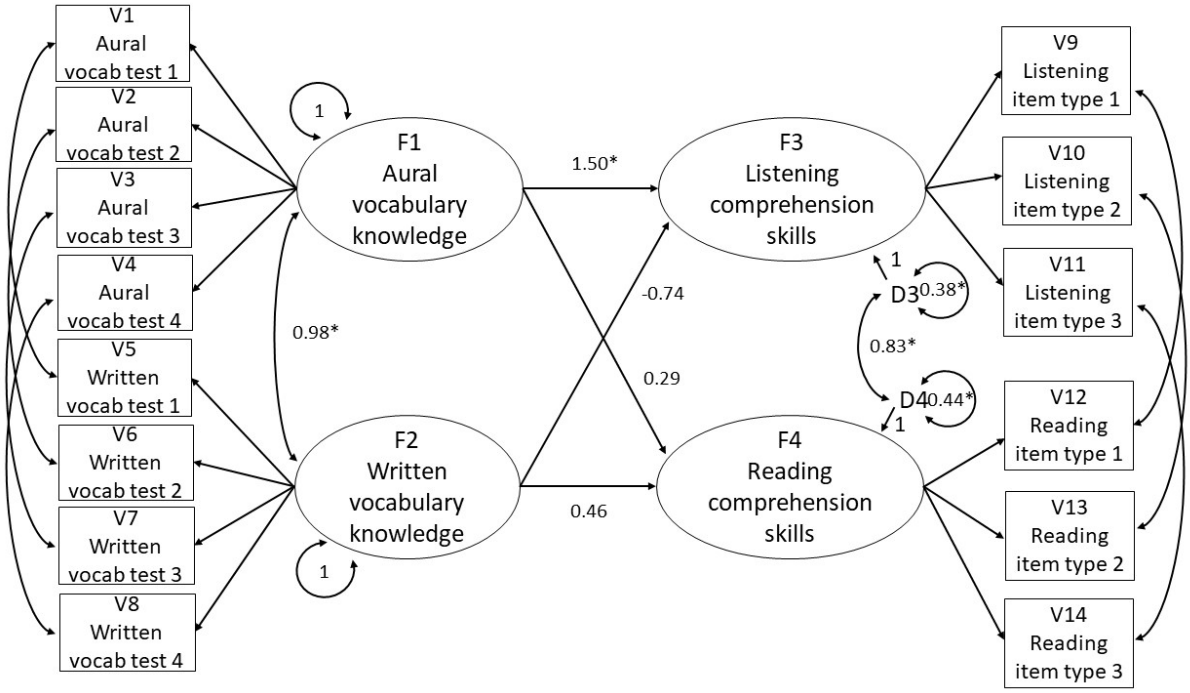


Figure 9. Standardized parameter estimates for the structural portion of the latent variable path model with the two-factor vocabulary knowledge structure. *$p < .05$.

Given the high correlation between the two predictors, a follow-up analysis was conducted with only a single predictor in the model at a time. The data-model fits of both single-predictor models were overall acceptable: $\chi^2 = 50.94$ ($df = 29$, $p = .007$), SRMR = .030, RMSEA = .064 with 90% CI [.033, .092], AIC = 14310.09, and CFI = .984 for the model with aural vocabulary and $\chi^2 = 49.99$ ($df = 29$, $p = .009$), SRMR = .034, RMSEA = .063 with 90% CI [.031, .091], AIC = 14164.68, and CFI = .985 for the model with written vocabulary. All the factors had strong loadings from their indicators with the standardized loadings 0.72 or higher in the model with aural vocabulary. Similarly, all the factors had strong loadings from their indicators with the standardized loadings 0.73 or higher in the model with written vocabulary.

The construct reliability measured by Coefficient *H* (Hancock & Mueller, 2001) for the model with aural vocabulary was .95 for the aural vocabulary knowledge factor, .85 for the listening comprehension skill factor, and .87 for the reading comprehension skill factor. The construct reliability for the model with written vocabulary was .96 for the written vocabulary knowledge factor, .85 for the listening comprehension skill factor, and .87 for the reading comprehension skill factor.

Figure 10 shows the standardized parameter estimates for the structural portions of both single-predictor models. All the structural paths were positive and statistically significant, suggesting that both aural and written vocabulary factors had the ability to explain both listening and reading skill factors. In the preceding analysis, where both predictors were included in the model, some of the structural paths were not significant. Given the usefulness of each vocabulary factor as a single predictor, the lack of significance of the paths in the preceding model could be considered due to the high correlation between the two predictors (which, e.g., increased the standard errors of the path coefficients).

It was also examined how much each of the vocabulary factors explained listening and reading skills as a single predictor. When only the aural vocabulary knowledge factor was included as a predictor, the factor explained 61% of the variance in the listening comprehension skill factor and 55% of the variance in the reading comprehension skill factor. When only the written vocabulary knowledge factor was included as a predictor, the factor explained 54% of the variance in the listening comprehension skill factor and 57% of the variance in the reading comprehension skill factor. In other words, the aural vocabulary knowledge explained the variance in listening better than the written vocabulary knowledge by 7%, and the written vocabulary knowledge explained the variance in reading better than the aural vocabulary

knowledge by 2%. Overall, these estimates seemed to be consistent with those in the analysis with both predictors included simultaneously.

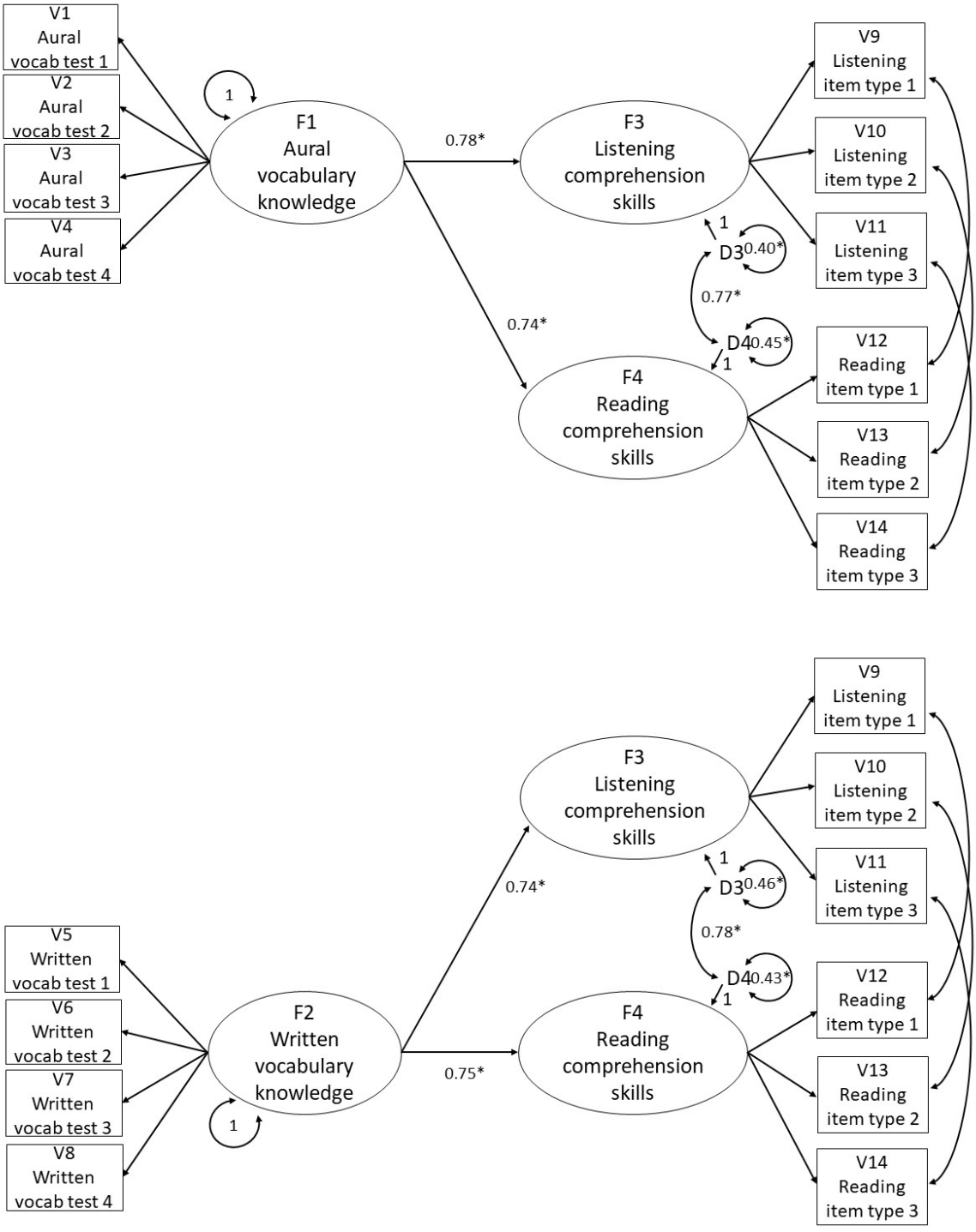


Figure 10. Standardized parameter estimates for the structural portions of the latent variable path models with the two-factor vocabulary knowledge structure with only a single predictor at a time. * $p < .05$.

Finally, latent variable path analysis was conducted with the one-factor vocabulary knowledge structure. Using Hu and Bentler's (1999) guidelines (i.e., acceptable when SRMR ≤ .08, RMSEA ≤ .06, and CFI ≥ .95), the data-model fit of the latent variable path model was overall acceptable, $\chi^2 = 117.20$ ($df = 67$, $p < .001$), SRMR = .033, RMSEA = .064 with 90% CI [.044, .082], AIC = 19526.59, and CFI = .980. All the factors had strong loadings from their indicators with the standardized loadings 0.72 or higher. The construct reliability measured by Coefficient *H* (Hancock & Mueller, 2001) was .97 for the vocabulary knowledge factor, .85 for the listening comprehension skill factor, and .87 for the reading comprehension skill factor.

Figure 11 shows the standardized parameter estimates for the structural portion of the model. The structural paths from the vocabulary knowledge factor to the listening and reading comprehension skill factors were positive and statistically significant, suggesting that the vocabulary factor had the ability to explain both listening and reading skill factors. The vocabulary factor explained 57% of the variance in the listening comprehension skill factor and 56% of the variance in the reading comprehension skill factor.

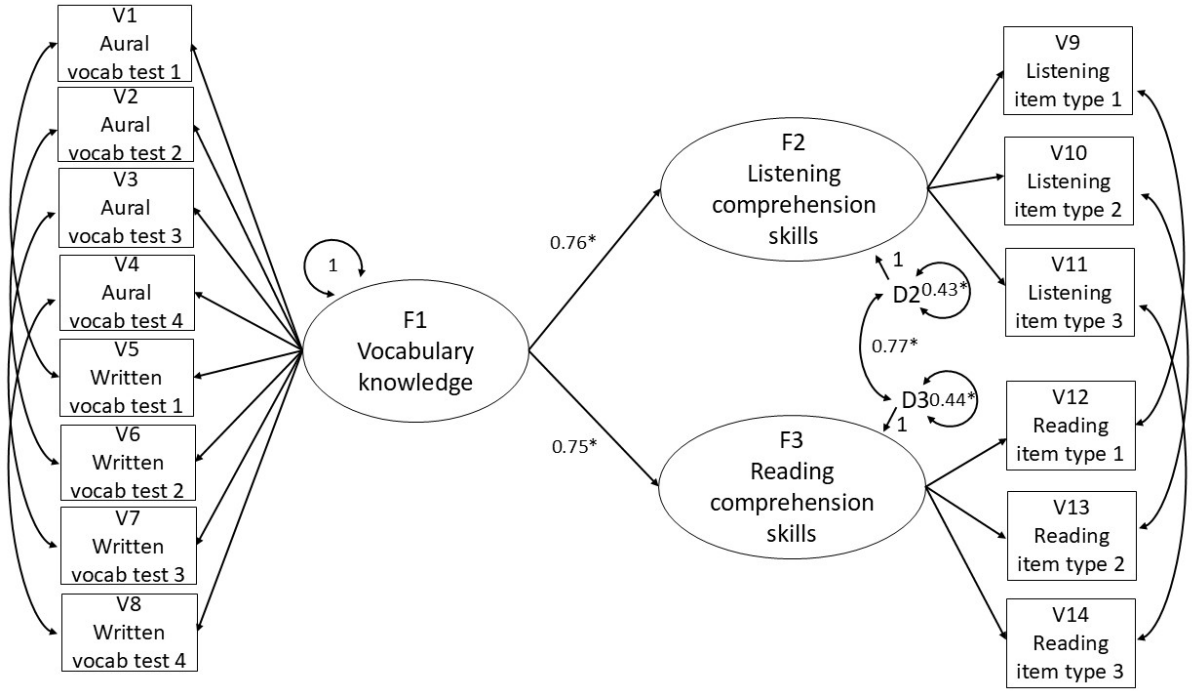


Figure 11. Standardized parameter estimates for the structural portion of the latent variable path model with the one-factor vocabulary knowledge structure. * $p < .05$.

Table 9 summarizes the proportions of variance in listening and reading comprehension skills explained by each of the vocabulary knowledge predictors. Overall, it seems that general vocabulary knowledge explained a large proportion of variance in both listening and reading comprehension skills, aural modality-specific knowledge explained a small proportion of variance in listening comprehension skills, and that written modality-specific knowledge explained virtually no proportion of variance in comprehension skills.

Table 9

Summary of the Proportions of Variance in Listening and Reading Comprehension Skills Explained by Each of the Vocabulary Knowledge Predictors

Vocabulary model	Predictor	Outcome	Variance explained
Two-factor model	Aural modality-specific knowledge	Listening	9%
	Written modality-specific knowledge	Listening	2%
	General vocabulary knowledge	Listening	51%
	Aural modality-specific knowledge	Reading	0%
	Written modality-specific knowledge	Reading	1%
	General vocabulary knowledge	Reading	55%
Two-factor model (with a single predictor at a time)	Aural vocabulary knowledge	Listening	61%
	Written vocabulary knowledge	Listening	54%
	Aural vocabulary knowledge	Reading	55%
	Written vocabulary knowledge	Reading	57%
One-factor model	General vocabulary knowledge	Listening	57%
	General vocabulary knowledge	Reading	56%

3.2.3 RQ3: The Relative Levels of Aural and Written Vocabulary Knowledge

The third research question asked whether the participants' aural vocabulary was larger, smaller, or equivalent to their written vocabulary. To this end, latent means modeling was conducted. Measurement invariance was examined stepwise, first for loadings and then for intercepts.

First, to examine loading invariance, each difference of the corresponding loadings between the aural and written vocabulary knowledge factors was coded as an additional parameter. The difference tests did not detect non-invariance for two pairs of loadings, *diff* = -1.41, *SE* = 0.80, *z* = -1.76, *p* = .079 for the form recall tests, and *diff* = -1.33, *SE* = 0.83, *z* = -1.60, *p* = .109 for the meaning recall tests. The other two pairs of loadings were non-invariant, *diff* = 2.62, *SE* = 1.11, *z* = 2.35, *p* = .019 for the yes-no tests, and *diff* = 2.74, *SE* = 0.62, *z* = 4.43, *p*

< .001 for the meaning selection tests. Therefore, loading constraints were imposed on the first two pairs of loadings but not on the other two pairs.

Next, intercept invariance was examined by assessing the model with intercept constraints. Note that intercept constraints were imposed only on the variables for which loading invariance held. The data-model fit of the constrained model was not very good, $\chi^2 = 58.99$ ($df = 17$, $p < .001$), SRMR = .041, RMSEA = .116 with 90% CI [.084, .148], and AIC = 11016.30. Modification indices indicated that the fit of the model could be significantly improved by releasing the intercept constraint for the meaning recall tests. After releasing the constraint, the data-model fit was overall acceptable, $\chi^2 = 29.74$ ($df = 16$, $p = .019$), SRMR = .017, RMSEA = .068 with 90% CI [.027, .106], and AIC = 10987.56. Although the data-model fit became acceptable, there was only one intercept constraint left in the model, which made it difficult to compare factor means. In fact, in this final model (Figure 12), where intercept constraint was retained only for the scale indicators, the estimate of the parameter a_{F2}, which was supposed to be a factor mean difference, turned out to be the same as the difference in means between the two scale indicators at the level of measured variables. This indicated that the attempt to assess factor means was not successful. Therefore, the relative levels of aural and written vocabulary knowledge were examined at the level of measured variables instead.

A paired-samples *t*-test was conducted to compare the mean of aural vocabulary test scores with that of written vocabulary for each of the four test formats. The results showed that the mean of aural vocabulary test scores was significantly lower than the mean of written vocabulary test scores with every test format: $t(183) = -10.78$, $p < .001$ for the form recall tests, $t(183) = -21.24$, $p < .001$ for the meaning recall tests, $t(180) = -19.48$, $p < .001$ for the yes-no tests, and $t(176) = -10.00$, $p < .001$ for the meaning selection tests. The effect size measured by

Cohen's *d* was 0.43 for the form recall tests, 0.79 for the meaning recall tests, 1.06 for the yes-no tests, and 0.47 for the meaning selection tests. Using the guideline of Plonsky and Oswald (2014) for within-group contrasts in L2 research (small when *d* was around 0.60, medium when *d* was around 1.00, and large when *d* was around 1.40), the effect size was considered small in the form recall tests and the meaning selection tests, small to medium in the meaning recall tests, and medium in the yes-no tests.

Also, the factor loadings of the latent means model were examined for an exploratory purpose. As seen in Figure 12, the meaning recall tests worked most effectively among the four test formats for both aural and written vocabulary knowledge factors. The meaning selection tests worked least effectively among the four test formats for both aural and written vocabulary knowledge factors. As mentioned earlier, there were two pairs of loadings that were non-invariant across the aural and written vocabulary factors, namely the loadings of the yes-no tests and the meaning selection tests. These tests worked less effectively for the written factor than for the aural factor.

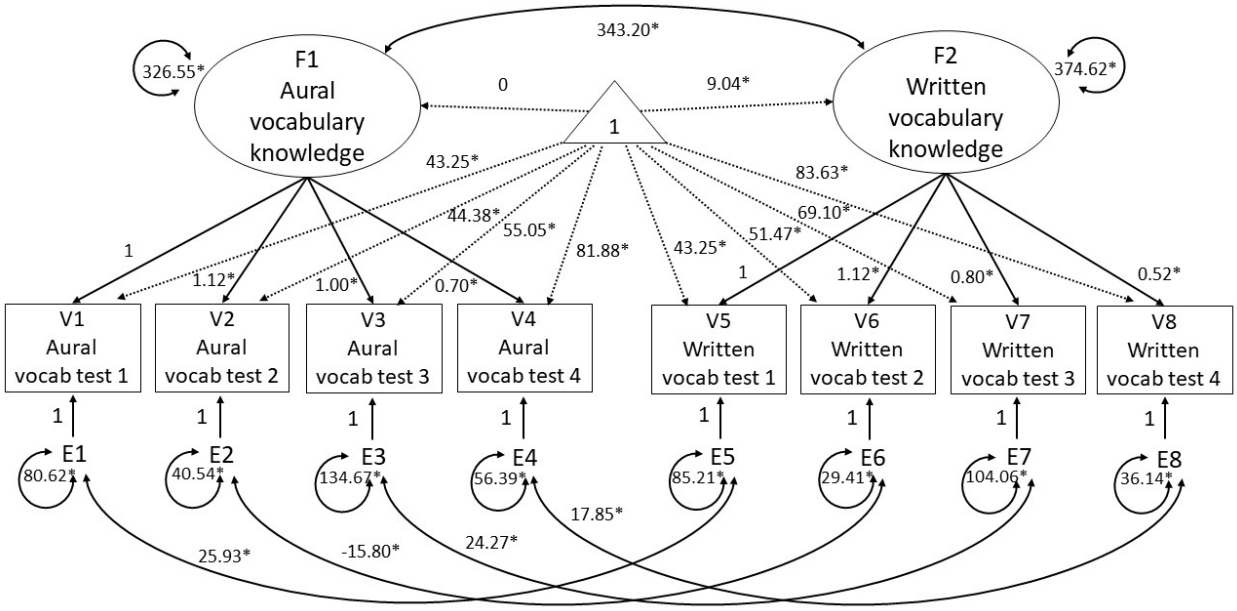


Figure 12. Unstandardized parameter estimates for the latent means model of aural and written vocabulary knowledge. *$p < .05$.

Chapter 4: Discussion

In the present study, the effect of modality (aural vs. written) was investigated in L2 vocabulary knowledge and in L2 comprehension skills. First, it was examined whether L2 learners' vocabulary knowledge was unidimensional or bi-dimensional in terms of the aural and written modes. It was then examined to what extent modality-specific aspects of vocabulary knowledge explained L2 listening and reading comprehension skills. The amount of aural and written vocabulary knowledge was also examined to see if there was any difference in the sizes of aural and written vocabulary knowledge. A latent modeling approach was taken in this study so that these research questions could be addressed without undue influence from measurement error and unique characteristics of particular tests. Furthermore, the current study was laid out to conduct a rigorous and robust investigation into modality effects; for fair comparison, parallel test formats were used across the modalities; to enhance comparability, the presentation time of target words was limited for both modalities (only once for aural and only 3 s for written); to minimize method effects, various formats of tests were used for each construct; to average out the effect of words, sets of target words were assigned to the tests across lists using Latin square design and counterbalanced across participants.

In the following sections, the findings of the present study will be discussed with each of the research questions in mind.

4.1 RQ1: The Structure of Vocabulary Knowledge

The first research question asked whether adult L2 learners' vocabulary knowledge was unidimensional or bi-dimensional in terms of the aural and written modalities. To this end, three confirmatory factor models were formed and compared. These models were: (1) one-factor

model with general vocabulary (no modality effect), (2) two-factor model with aural and written vocabulary, and (3) residualized-factor model with general vocabulary and aural and written modality-specific skills. The results of data-model fits showed that, whereas the residualized-factor model did not fit the data well and resulted in an improper solution, the one-factor model and the two-factor model demonstrated overall acceptable data-model fits. A Satorra-Bentler scaled chi-square difference test further showed that the two-factor model was better than the one-factor model and the difference was statistically significant. However, when the two-factor model was examined, the factor correlation turned out to be very high (r = 0.98). Although the two-factor model was better statistically, the aural and written vocabulary knowledge factors practically overlapped to a considerable degree.

This finding is actually similar to those of previous studies with SEM that examined other aspects of vocabulary knowledge such as size versus depth (González-Fernández, 2022; González-Fernández & Schmitt, 2020; Koizumi & In'nami, 2020), recall versus recognition (Stewart et al., 2024), and time-sensitive versus untimed (Hui et al., 2022). Among these studies, some showed a significantly better fit for a multi-factor model over a one-factor model as with the present study (Koizumi & In'nami, 2020; Stewart et al., 2024), while others did not show such a difference (González-Fernández, 2022; González-Fernández & Schmitt, 2020; Hui et al., 2022), but importantly, they all showed a high factor correlation (e.g., r = 0.92 in Hui et al., 2022; r = 0.95 in Koizumi & In'nami, 2020; r = 0.85 in Stewart et al., 2024). This appears to suggest that the impact of general vocabulary knowledge is very strong whatever aspect of vocabulary knowledge is considered. The finding of the present study indicated that the aspect of aural versus written modality was no exception.

This finding is a little different from that of the study by Cheng and Matthews (2018), one of the most closely related studies to the current investigation. Using an exploratory factor analysis, they clearly showed a two-factor solution for their vocabulary tests, where a written meaning recognition test and a written form recall test loaded onto one factor, and an aural form recall test loaded onto another factor, thereby suggesting (somewhat more clearly than the present study) that there is a dimension of aural versus written modality in vocabulary knowledge. This difference might have resulted from the differences in test formats. Cheng and Matthews (2018) used a partial dictation test for their aural vocabulary test, which required participants to identify aural target words in sentences. This presumably required segmentation skills on top of simple aural vocabulary knowledge. The present study, on the other hand, presented target words in isolation. Therefore, based on the present study, it might be concluded that the impact of modality effects is limited when words are presented in isolation. As the study by Cheng and Matthews (2018) was exploratory and included only one format of aural vocabulary test, further research is needed to examine whether the limited modality effect extends to sentence contexts or not.

4.2 RQ2: The Unique Contribution of Modality to Comprehension

The second research question asked the extent to which the modality-specific aspects of L2 vocabulary knowledge explained L2 listening and reading comprehension skills. To this end, latent variable path analysis was conducted with the two-factor model of vocabulary knowledge explaining the factors of listening and reading comprehension skills. This path model had an acceptable data-model fit. To estimate the unique contribution of aural vocabulary knowledge to listening comprehension above and beyond what written vocabulary knowledge could explain, delta R^2 was calculated, that is, the variance in listening comprehension skills explained by the

two vocabulary factors minus the variance explained by the written vocabulary factor alone. The delta R^2 was 0.09, indicating that the aural vocabulary knowledge factor uniquely explained 9% of the variance in the listening comprehension skill factor. Using the same method, the unique contribution of written vocabulary knowledge to reading comprehension was also estimated, but the unique contribution was virtually none (1%). In sum, aural vocabulary knowledge uniquely explained some variance in listening comprehension skills, but written vocabulary knowledge did not uniquely explain variance in reading comprehension skills.

These findings—modality effects in the aural but not in the written modality—have been observed in a study by Uchihara (2023) also. Sampling participants from the same L1 Japanese learner population as the present study, the researcher examined the effect of modality of weekly vocabulary quizzes in a college English class context. In that study, participants in one group studied for aural vocabulary quizzes while participants in another group studied for written vocabulary quizzes over 10 weeks. The results of pre- and post-tests of aural and written vocabulary knowledge showed that the participants in the aural test group improved in aural vocabulary knowledge significantly more than those in the written test group while there was no group difference in written vocabulary knowledge. Why were modality effects evident in the aural mode but not in the written mode for the study by Uchihara (2023) as well as the current one? One possibility is that, for adult L1 Japanese learners, written vocabulary knowledge almost always serves as the foundation of their vocabulary knowledge, and aural vocabulary knowledge tends to be a subsidiary part of their vocabulary knowledge. In other words, when these L2 learners have aural knowledge of words, they almost always have the corresponding written knowledge (but not vice versa), and therefore written vocabulary knowledge and general vocabulary knowledge basically overlap and do not produce any unique effect of written

modality. Aural vocabulary knowledge, however, is in a sense an extra part and might produce some unique effects. In fact, when the average aural and written vocabulary test scores were compared for each of the current 185 participants, the written vocabulary test scores were higher than the aural test scores for as many as 181 participants (98%), which suggests written vocabulary knowledge as the foundation of their vocabulary knowledge, although this speculation needs to be confirmed with longitudinal studies in the future. It is also important to stress that these findings for L1 Japanese learners should not be generalized to other learner populations without replication. However, given that adult L2 learners tend to better learn explicitly than implicitly (DeKeyser & Larson-Hall, 2005) and written language is conducive to explicit learning, I suspect a similar pattern of results for other adult L2 learner populations, which, again, needs to be tested in future studies, though. From a pedagogical point of view, the findings might suggest it would be useful for L2 learners and teachers to pay attention to the weaker aspect of knowledge (e.g., aural vocabulary in the current case) in order to stretch their learning edges.

Although aural vocabulary knowledge uniquely contributed to listening comprehension, the amount of contribution seems to have been small relative to that of general vocabulary knowledge. While the aural vocabulary knowledge factor uniquely explained 9% of the variance in the listening comprehension skill factor, the common part of the two vocabulary factors—general vocabulary knowledge—explained 51% of the variance in the listening comprehension skill factor. As for reading, the common part of the two vocabulary factors explained 55% of the variance in the reading comprehension skill factor with virtually no contribution from modality-specific knowledge. These results suggest that, although a modality effect does seem to exist, the impact of general vocabulary knowledge is much larger in L2 comprehension. That said,

modality effects need further investigation as other recent studies suggested a larger effect of modality when aural vocabulary was presented in sentence contexts as opposed to in isolation. Masrai (2022), for example, used two aural vocabulary tests, an aural yes-no test and a partial dictation test, and found a stronger correlation of a listening comprehension test with the dictation test ($r = 0.74$) than with the yes-no test ($r = 0.59$), suggesting, among other things, the potential importance of identifying aural words in contexts for listening comprehension. Similarly, Saito et al. (2023) included two types of aural vocabulary tests, a multiple-choice meaning selection test and a lexicosemantic judgment test. While target words were presented in isolation in the former test, target words were presented in sentence contexts in the latter test and the participants were required to make judgments on semantic appropriateness (e.g., *My grandfather bought an estate* vs. *My friend's estate was very kind*). The researchers found a stronger correlation of a listening comprehension test with the lexicosemantic judgment test ($r = 0.66$) than with the meaning selection test ($r = 0.43$). Although the effect of modality was limited relative to general vocabulary knowledge for comprehension in the present study, it was based on the vocabulary tests that presented words in isolation, and so further research is needed to examine modality effects in vocabulary knowledge in sentence contexts in relation to comprehension. Preferably, such investigation should be conducted at the level of latent constructs as in the present study so that the findings will be less affected by measurement error and unique characteristics of particular tests.

4.3 RQ3: The Relative Levels of Aural and Written Vocabulary Knowledge

The third research question asked whether the participants' aural vocabulary was larger, smaller, or equivalent to their written vocabulary. To this end, latent means modeling was conducted. Unfortunately, however, measurement invariance could not be established for this

model; when loading and intercept invariance was examined stepwise, two out of four pairs of loadings (i.e., yes-no and meaning selection tests) were found to be non-invariant, and three out of four pairs of intercepts (i.e., yes-no, meaning selection, and meaning recall tests) were found to be non-invariant, which did not allow for comparing means at the latent level. It is hard to determine the cause of this result, but it may be that many formats of tests function differently across the constructs of aural and written vocabulary knowledge and the present study was not able to include a large enough number of tests that measure these constructs comparably. For instance, the yes-no tests and the meaning selection tests worked less effectively for the written vocabulary construct than for the aural vocabulary construct in this study. These multiple-choice tests would have been more likely to be affected by the test-takers' test-wise strategies compared with recall tests and it might have been the case that such test-wise strategies were more easily utilized in the written tests than in the aural tests, possibly resulting in the less effectiveness of the written multiple-choice tests. Future studies should try latent mean comparison with improved test instruments and with a greater number of indicators.

Due to the measurement non-invariance, the participants' aural and written vocabulary knowledge was compared at the level of measured variables. The results showed that the participants' aural test scores were significantly lower than their written test scores for all the four vocabulary tests, suggesting that the participants had smaller aural vocabulary than written vocabulary. This finding is in line with those of previous studies with the same L1 Japanese learner population (Hamada & Yanagawa, 2024; Mizumoto & Shimamoto, 2008) as well as other L1 populations (e.g., Chinese, Milton et al., 2010; Greek, Milton & Hopkins, 2006). However, the finding should be further replicated with more populations with different types of learners to examine generalizability.

It is also interesting to note the effect size differences across different test formats. The score difference between aural and written tests was relatively small with the form recall tests (d = 0.43) and with the meaning selection tests (d = 0.47), and it was larger with the meaning recall tests (d = 0.79) and with the yes-no tests (d = 1.06). The latter two test formats produced larger differences between aural and written tests perhaps because in the tests the participants did not have control over the presentation of target words, unlike in the form recall tests, and the participants did not have support when they provided answers, unlike in the meaning selection tests. It is particularly interesting to see that the meaning recall tests and the meaning selection tests—although they shared all features except for the type of answer response—produced very different effect sizes (d = 0.79 vs. d = 0.47). The support from answer options reduced the difference between aural and written tests in the meaning selection tests. These observations suggest that researchers and teachers should make conscious choices of test formats taking their test purpose into consideration.

Also, the factor loadings of the latent means model were examined for an exploratory purpose. The meaning recall tests worked most effectively among the four test formats to capture the constructs of vocabulary knowledge, and the meaning selection tests worked least effectively. This was the case for both aural and written constructs. These results are in agreement with the research work by Stuart McLean and colleagues, who have argued for the superiority of meaning recall tests over meaning recognition tests based, for example, on their higher internal consistency and greater predictive validity for comprehension skills (e.g., Komiya & McLean, 2024; McLean et al., 2020; Stewart et al., 2024).

4.4 Limitations and Future Directions

As with any study, the present study has limitations. These limitations need to be kept in mind when interpreting the findings. The limitations also indicate future directions on how to improve and expand this line of investigation.

As for test reliability, the current vocabulary tests had good internal consistency with Cronbach's alpha estimates ranging from .82 to .94. Using the guideline of George and Mallery (2003), these values are in the range of "good" or "excellent." The listening and reading comprehension tests also had good internal consistency (.85 and .86, respectively). However, the internal consistency of nonword test items in the yes-no tests and of comprehension test items per item type was not so good with Cronbach's alpha estimates ranging from .60 to .70, which are in the range of "questionable." This difference in internal consistency was probably caused by the difference in the number of items; in this study when the number of items was 40 or higher, they had high internal consistency, but when the number of items was 20 or lower, they had lower internal consistency. These results suggest that future studies should include at least 40 items per construct to have reliable measures (see also McLean et al., 2020, who examined the internal consistency of various types of vocabulary tests with difference numbers of items and found 40 to be the number required to reach internal consistency higher than .80 across various types of tests).

Limited generalizability is also important to note here. As already mentioned several times, the current sample was drawn from adult L1 Japanese learners of English in Japan with little or no overseas experience. These learners often tend to have a learning background with greater focus on written than aural language (Kobayashi, 2001), which may or may not be true for other populations. Many other factors, including, but not limited to, age, L1, and context,

may also change the current findings. Therefore, the present study should be replicated with different populations. For instance, it might be interesting to examine L1 Arabic learners as they previously showed the opposite tendency to the present findings (i.e., larger aural vocabulary than written; Milton & Hopkins, 2006). It is also important to examine child L2 learners as well as immigrant learners. Also, for generalizability, languages other than English need to be investigated as the vast majority of studies, including the current one, have been using English as a target language in this line of research (but see, e.g., Robles-García et al., 2024 for the development of L2 Spanish Vocabulary Levels Test).

As noted earlier, all vocabulary tests in the present study tested knowledge of words in isolation, which is also an important aspect to consider. Although the effects of modality were somewhat limited in this study, the effects could become larger if words are present in sentence contexts (see, e.g., Masrai, 2022; Saito et al., 2023). This possibility should be examined in future studies, preferably with a latent variable framework as in the present study, in order to estimate modality effects for knowledge of words in sentence contexts versus in isolation.

As an extension of this study, it would also be interesting to examine grammar knowledge. As previous studies showed that L2 learners performed differently between aural and written grammaticality judgment tests (e.g., Granena, 2013; Johnson & Newport, 1989 vs. Johnson, 1992), modeling aural and written L2 grammar knowledge at the latent level seems to be a promising endeavor for future research.

The limitations of the listening comprehension test should also be mentioned here. The present study used the TOEIC listening test items, which were based on scripted speech. As scripted speech is qualitatively different from naturally occurring speech (Biber et al., 2004;

Redeker, 1984), further work is needed to explore the extent to which listening skills assessed with scripted speech may be generalized to listening skills in the real world. A follow-up study to the current one could also include more spontaneous speech for listening test materials (see, e.g., Clark, 2014 for the use of semi-scripted speech for a listening comprehension test). It is very possible that having such a more realistic listening test as an outcome variable will make the effect of modality stronger.

The selection of target words also potentially has room for improvement. Spoken language in listening comprehension tests often mostly consists of high frequency, non-academic words. A vocabulary test tailored for the purpose of listening comprehension would close the gap in words between vocabulary tests and listening comprehension tests and improve predictive power, which might be an interesting topic for future research.

It is also important to note that this study regarded knowledge of form-meaning mapping in each modality as a set of skills and did not try to locate where problems occurred within the set of skills. The present study, therefore, could not tell, for example, why the participants' aural vocabulary test scores were lower than their written scores (they might have failed to discriminate one phoneme from another, their lexical representations might have been fuzzy, aural forms might not have been registered in their mental lexicon in the first place, etc.; see, e.g., Gor, 2015 for a review of such finer-grained issues). Further research is needed to better understand the issues around modality effects. For example, it would be useful for future studies to include a sound discrimination test to assess the impact of lower-level processing ability (for such an instrument, see, e.g., Chrabaszcz & Gor, 2014; Wilson et al., 2011; Wong et al., 2017).

4.5 Conclusion

In the present study, the effects of aural versus written modality were examined in relation to L2 vocabulary knowledge and comprehension skills. It was first examined whether L2 vocabulary knowledge was unidimensional or bi-dimensional in terms of the aural and written modes. It was then investigated how these modality-specific aspects contributed to L2 listening and reading comprehension. The study also examined the participants' relative levels of aural and written vocabulary knowledge. Latent variable statistical methods were used to minimize influence from measurement error and unique characteristics of particular tests. Confounding factors were further mitigated by using parallel vocabulary tests across modalities and counterbalancing target words across participants. The results of this study were nuanced. Modality effects were observed in the sense that (1) the two-factor model of vocabulary knowledge had a significantly better fit to the data than the one-factor model, (2) aural vocabulary knowledge uniquely explained 9% of the variance in listening comprehension skills, and (3) the participants' aural vocabulary size was significantly smaller than their written vocabulary size. However, the effects of modality were limited in the sense that (1) the aural and written vocabulary knowledge factors were very highly correlated ($r = 0.98$) and (2) general vocabulary knowledge explained much more variance in comprehension skills than modality-specific knowledge (51% vs. 9% for listening and 55% vs. 1% for reading). These results suggest that, although aural versus written modality effects do seem to exist in L2 vocabulary knowledge and comprehension skills, its practical impact is small compared with that of general vocabulary knowledge at least in the context where words are presented in isolation as in the present study. Further studies are needed to examine the generalizability of the current findings to other populations, languages, and contexts.

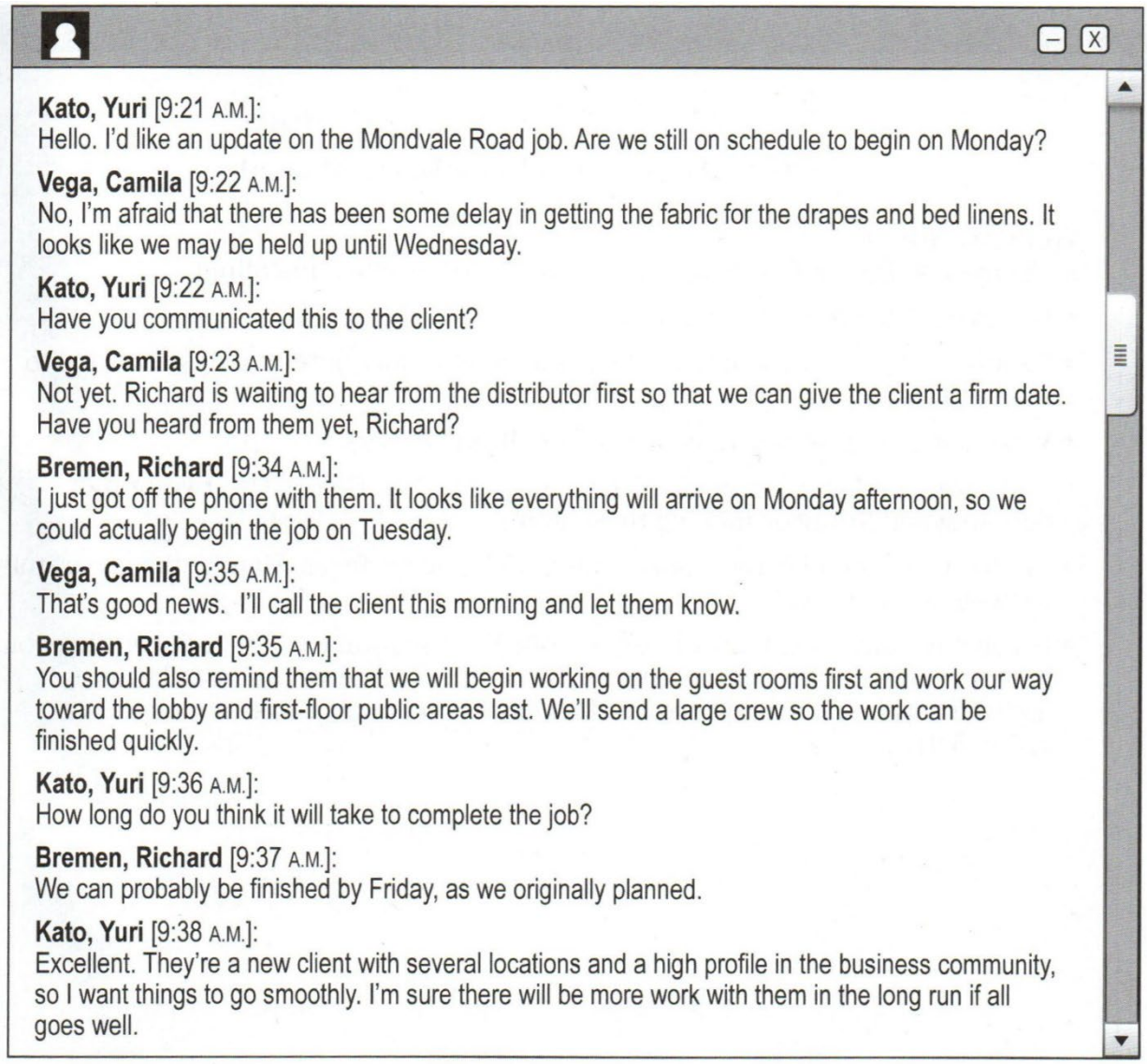
Appendices

Appendix A The Participants' College Majors

Major	Count	Percentage
Economics	19	10%
Law	16	9%
Computer science	13	7%
English language, literature, and culture	13	7%
Japanese language, literature, and culture	13	7%
Engineering	11	6%
Biology	10	5%
Psychology	8	4%
Agriculture	6	3%
Chemistry	6	3%
International communication	6	3%
Education	5	3%
Liberal arts	5	3%
Linguistics	5	3%
Sociology	4	2%
Medicine	3	2%
Nursing	3	2%
Pharmacy	3	2%
Polish	3	2%
Politics	3	2%
Dentistry	2	1%
Materials science	2	1%
Math	2	1%
Nutrition	2	1%
Tourism	2	1%
Others (n = 1): Anthropology, Archaeology, Architecture, Classics, Dance, Environmental science, Fine arts, French literature, Geography, History, Home economics, Journalism, Marine science, Mongolian, Philosophy, Urban planning, Urdu, Veterinary medicine	18	10%
No college degree	2	1%

Appendix B Examples of Reading Test Items

Text:



Kato, Yuri [9:21 A.M.]:
Hello. I'd like an update on the Mondvale Road job. Are we still on schedule to begin on Monday?

Vega, Camila [9:22 A.M.]:
No, I'm afraid that there has been some delay in getting the fabric for the drapes and bed linens. It looks like we may be held up until Wednesday.

Kato, Yuri [9:22 A.M.]:
Have you communicated this to the client?

Vega, Camila [9:23 A.M.]:
Not yet. Richard is waiting to hear from the distributor first so that we can give the client a firm date. Have you heard from them yet, Richard?

Bremen, Richard [9:34 A.M.]:
I just got off the phone with them. It looks like everything will arrive on Monday afternoon, so we could actually begin the job on Tuesday.

Vega, Camila [9:35 A.M.]:
That's good news. I'll call the client this morning and let them know.

Bremen, Richard [9:35 A.M.]:
You should also remind them that we will begin working on the guest rooms first and work our way toward the lobby and first-floor public areas last. We'll send a large crew so the work can be finished quickly.

Kato, Yuri [9:36 A.M.]:
How long do you think it will take to complete the job?

Bremen, Richard [9:37 A.M.]:
We can probably be finished by Friday, as we originally planned.

Kato, Yuri [9:38 A.M.]:
Excellent. They're a new client with several locations and a high profile in the business community, so I want things to go smoothly. I'm sure there will be more work with them in the long run if all goes well.

Question 1: What kind of business does the client most likely own?

(A) A shipping company

(B) A fabric manufacturing factory

(C) A hotel chain

(D) A design firm

Question 2: When will the crew begin work?

(A) On Monday

(B) On Tuesday

(C) On Wednesday

(D) On Friday

Question 3: What will Ms. Vega most likely do next?

(A) Deliver a shipment of drapes

(B) Organize a large work crew

(C) Call the fabric distributor

(D) Contact the client

Question 4: At 9:38 A.M., what does Ms. Kato mean when she writes, “in the long run”?

(A) She is pleased that the client is located nearby.

(B) She is proud of her company’s history of high-quality performance.

(C) She believes that the work will be more expensive than expected.

(D) She thinks that there could be additional work with the client in the future.

Appendix C Examples of Listening Test Items

Script: Hi, Li Na. This is Youssef. I'm planning the luncheon to celebrate Nancy Baxter's retirement from our marketing department. I went to the store yesterday to purchase her favorite chocolates for us to give to her at the luncheon. Now, we need someone who can speak about some highlights of her career. Of course, many people have worked with Nancy on successful ad campaigns over the years, but you've worked with her the most.

Question 1: What department does the speaker work in?

(A) Human resources

(B) Customer service

(C) Research

(D) Marketing

Question 2: What does the speaker say he did yesterday?

(A) He bought a gift.

(B) He sent an invitation.

(C) He contacted a caterer.

(D) He spoke with a client.

Question 3: What does the speaker imply when he says, “you've worked with her the most”?

(A) The listener may be promoted.

(B) A cost should be shared.

(C) The listener should give a speech.

(D) A staffing decision was unfair.

Note that, following the format of the TOEIC test, the script and the questions (but not the answer options) were aurally presented. The questions and the answer options (but, of course, not the script) were presented on the computer screen.

Appendix D Translation of Background Questionnaire

Participant ID: _____

Age: _____

Gender: _____

Occupation: _____

Major at College: _____

Please make sure that you meet ALL the following eligibility criteria.

- □ Between the ages of 18 to 40
- □ Native speaker of Japanese
- □ Born and raised in a monolingual Japanese-speaking family
- □ The medium of education was Japanese at least until high school graduation
- □ No experience staying abroad for more than a month

Appendix E The Results of the Analysis of Vocabulary Knowledge Structure With Z-scores

It was examined what type of vocabulary knowledge structure would best fit the data using z-scores.

Model 1, a one-factor structure with general vocabulary, had the following results of data-model fit: $\chi^2 = 45.63$ ($df = 16$, $p < .001$), SRMR = .021, RMSEA = .100 with 90% CI [.067, .135], AIC = 2424.51, and CFI = .979. Using Hu and Bentler's (1999) guidelines (i.e., acceptable when SRMR ≤ .08, RMSEA ≤ .06, and CFI ≥ .95), the SRMR and CFI values were acceptable, whereas the RMSEA value was not in the recommended range. The vocabulary factor had strong loadings from the indicators with the loadings 0.72 or higher.

Model 2, a two-factor structure with aural and written vocabulary, had the following results of data-model fit: $\chi^2 = 29.44$ ($df = 15$, $p = .014$), SRMR = .020, RMSEA = .072 with 90% CI [.032, .110], AIC = 2408.75, and CFI = .990. As with Model 1, the SRMR and CFI values were acceptable. The RMSEA value was better than that of Model 1, but it was slightly above the recommended range. Both aural and written vocabulary factors had strong loadings from their indicators with the loadings 0.73 or higher. The two factors were highly correlated ($r = 0.97$), indicating that the two factors overlapped to a considerable degree.

Model 3, a residualized-factor structure with general vocabulary and aural and written modality-specific skills, resulted in an improper solution with the latent variable covariance matrix not being positive definite with a negative variance.

To compare Model 1 and Model 2, a Satorra-Bentler scaled chi-square difference test was conducted. The difference was statistically significant, $\Delta\chi^2_{(SB)} = 14.01$ ($df = 1$, $p < .001$), suggesting that Model 2 had a significantly better fit to the data than Model 1. Model 2 looked better than Model 1 based on other model fit indices also.

Appendix F Scoring Guidelines for the Form Recall Tests

- Mark it correct even if the response includes minor misspelling/mispronunciation as long as it is recognized as the target word and it is not confused with another word
- Mark it incorrect if the response does not start with the initial letter(s)/phoneme(s) provided
- Mark it correct if the response is a synonym of the target word as long as it starts with the initial letter(s)/phoneme(s) provided
- Mark it correct even if the response is in a different part of speech (e.g., *honestly* for the target word *honest*)
- Mark it incorrect if the response lacks some meaning of the target word (e.g., *desire* for the target word *desirable*) or has some extra meaning (e.g., *poet* for the target word *poem*)

Appendix G Scoring Guidelines for the Meaning Recall Tests

- Mark it correct even if the response includes minor L1 issues (e.g., using wrong *kanji*)
- Mark it correct even if the response is not the expected L1 word as long as it is acceptable as a meaning of the target English word
- Mark it correct even if the response is in a different part of speech (e.g., 出発する for the target word *departure*)
- Mark it incorrect if the response lacks some meaning of the target word (e.g., 蜂 for the target word *honey*) or has some extra meaning (e.g., 航海士 for the target word *voyage*)
- Mark the response incorrect if the participant just typed the target English word in *katakana* (e.g., レジェンド for the target word *legend*)

Appendix H The Results of the Analysis of Vocabulary Knowledge Structure With Loading Constraints

It was examined what type of vocabulary knowledge structure would best fit the data with potential loading constraints imposed on parallel test formats.

First, loading invariance was examined for Model 1, a one-factor structure with general vocabulary. The difference tests did not detect non-invariance for two pairs of loadings, *diff* = -1.50, *SE* = 0.80, *z* = -1.87, *p* = .062 for the form recall tests, and *diff* = -1.32, *SE* = 0.83, *z* = -1.59, *p* = .111 for the meaning recall tests. The other two pairs of loadings were non-invariant, *diff* = 2.50, *SE* = 1.11, *z* = 2.26, *p* = .024 for the yes-no tests, and *diff* = 2.78, *SE* = 0.62, *z* = 4.48, *p* < .001 for the meaning selection tests. Therefore, loading constraints were imposed on the first two pairs of loadings but not on the other two pairs. With the constraints, Model 1 had the following results of data-model fit: χ^2 = 42.25 (*df* = 18, *p* = .001), SRMR = .032, RMSEA = .085 with 90% CI [.052, .119], AIC = 10997.10, and CFI = .985. Using Hu and Bentler's (1999) guidelines (i.e., acceptable when SRMR ≤ .08, RMSEA ≤ .06, and CFI ≥ .95), the SRMR and CFI values were acceptable, whereas the RMSEA value was not in the recommended range. The vocabulary factor had strong loadings from the indicators with the loadings 0.83 or higher.

Next, loading invariance was examined for Model 2, a two-factor structure with aural and written vocabulary. The difference tests did not detect non-invariance for two pairs of loadings, *diff* = -1.41, *SE* = 0.80, *z* = -1.76, *p* = .079 for the form recall tests, and *diff* = -1.33, *SE* = 0.83, *z* = -1.60, *p* = .109 for the meaning recall tests. The other two pairs of loadings were non-invariant, *diff* = 2.62, *SE* = 1.11, *z* = 2.35, *p* = .019 for the yes-no tests, and *diff* = 2.74, *SE* = 0.62, *z* = 4.43, *p* < .001 for the meaning selection tests. Therefore, loading constraints were imposed on the first two pairs of loadings but not on the other two pairs. With the constraints, Model 2 had the

following results of data-model fit: $\chi^2 = 29.74$ ($df = 16$, $p = .019$), SRMR = .017, RMSEA = .068 with 90% CI [.027, .106], AIC = 10987.56, and CFI = .991. As with Model 1, the SRMR and CFI values were acceptable. The RMSEA value was better than that of Model 1, but it was slightly above the recommended range. Both aural and written vocabulary factors had strong loadings from their indicators with the loadings 0.84 or higher. The two factors were highly correlated ($r = 0.98$), indicating that the two factors overlapped to a considerable degree.

Model 3, a residualized-factor structure with general vocabulary and aural and written modality-specific skills, had an issue of non-convergence and did not work.

To compare Model 1 and Model 2, a Satorra-Bentler scaled chi-square difference test was conducted. The difference was statistically significant, $\Delta\chi^2_{(SB)} = 13.31$ ($df = 2$, $p = .001$), suggesting that Model 2 had a significantly better fit to the data than Model 1. Model 2 looked better than Model 1 based on other model fit indices also.

Appendix I Derivation of Delta R^2 Using the Path Tracing Rules

Standardized path tracing rules:

In a given trace…

- One can go forward or backward causally; but once gone forward, one cannot go backward.
- One can go through only one unanalyzed relation (two-headed arrow).
- One can enter a variable no more than once; one can leave a variable no more than once.

The proportion of variance in F3 (listening skills) explained by F1 (aural vocabulary) and F2 (written vocabulary) taken together is the sum of the paths from F3 back to itself, that is, using the path tracing rules:

$$\beta^2_{F3F2} + \beta^2_{F3F1} + 2r_{F2,\,F1}\beta_{F3F1}\beta_{F3F2}$$

The proportion of variance in F3 (listening skills) explained by F2 (written vocabulary) alone is $r^2_{F2,\,F3}$, that is, using the path tracing rules:

$$(\beta_{F3F2} + r_{F2,\,F1}\beta_{F3F1})^2 = \beta^2_{F3F2} + r^2_{F2,\,F1}\beta^2_{F3F1} + 2r_{F2,\,F1}\beta_{F3F1}\beta_{F3F2}$$

The difference between the two above is the delta R^2 for F1 (aural vocabulary), that is:

$$(\beta^2_{F3F2} + \beta^2_{F3F1} + 2r_{F2,\,F1}\beta_{F3F1}\beta_{F3F2}) - (\beta^2_{F3F2} + r^2_{F2,\,F1}\beta^2_{F3F1} + 2r_{F2,\,F1}\beta_{F3F1}\beta_{F3F2})$$

$$= \beta^2_{F3F1} - r^2_{F2,\,F1}\beta^2_{F3F1}$$

$$= \beta^2_{F3F1}(1 - r^2_{F2,\,F1})$$

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*(S2), 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* (RM-04-03, TOEFL-MS-25). https://www.ets.org/Media/Research/pdf/RM-04-03.pdf

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego, CA: Academic Press.

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, *14*(1), 14–31. https://doi.org/10.1080/10904018.2000.10499033

Brysbaert, M., Keuleers, E., & Mandera, P. (2021). Which words do English non-native speakers know? New supernational levels based on yes/no decision. *Second Language Research*, *37*(2), 207–231. https://doi.org/10.1177/0267658320934526

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

Cheng, J., Matthews, J., Lange, K., & McLean, S. (2023). Aural single-word and aural phrasal verb knowledge and their relationships to L2 listening comprehension. *TESOL Quarterly*, *57*(1), 213–241. https://doi.org/10.1002/tesq.3137

Chrabaszcz, A., & Gor, K. (2014). Context effects in the processing of phonolexical ambiguity in L2. *Language Learning*, *64*(3), 415–455. https://doi.org/10.1111/lang.12063

Clark, M. (2014). The use of semi-scripted speech in a listening placement test for university students. *Papers in Language Testing and Assessment*, *3*(2), 1–26.

Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *The Journal of Educational Research*, *36*(3), 206–217. https://doi.org/10.1080/00220671.1942.10881160

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284. https://doi.org/10.1016/j.wocn.2005.06.002

Deegan, J., Jr. (1978). On the occurrence of standardized regression coefficients greater than one. *Educational and Psychological Measurement*, *38*(4), 873–888. https://doi.org/10.1177/001316447803800404

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean?. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). New York: Oxford University Press.

Field, J. (2013). Cognitive validity. In A. Geranpaye & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge: Cambridge University Press.

Geiser, C. (2012). *Data analysis with Mplus*. New York: Guilford Publications.

Geiser, C. (2022, June 7). *Standardized path coefficients larger than 1 ???* [Video]. YouTube. https://www.youtube.com/watch?v=R_kILwAHFuo&ab_channel=QuantFish

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference 11.0 Update* (4th ed.). Boston: Allyn & Bacon.

González-Fernández, B. (2022). Conceptualizing L2 vocabulary knowledge: An empirical examination of the dimensionality of word knowledge. *Studies in Second Language Acquisition*, *44*(4), 1124–1154. https://doi.org/10.1017/S0272263121000930

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

Gor, K. (2015). Phonology and morphology in lexical processing. In J. W. Schwieter (Ed.), *The Cambridge handbook of bilingual processing* (pp. 173–199). Cambridge: Cambridge University Press.

Granena, G. (2013). Reexamining the robustness of aptitude in second language acquisition. In G. Granena, & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 179–204). Amsterdam: Benjamins.

Ha, H. T. (2021). Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. *Language Testing in Asia*, *11*(1), 1–20. https://doi.org/10.1186/s40468-021-00131-8

Hamada, Y., & Yanagawa, K. (2024). Aural vocabulary, orthographic vocabulary, and listening comprehension. *International Review of Applied Linguistics in Language Teaching*, *62*(2), 953–975. https://doi.org/10.1515/iral-2022-0100

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*(3), 373–388. https://doi.org/10.1007/BF02294440

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future — A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Hancock, G. R., & Schoonen, R. (2015). Structural Equation Modeling: Possibilities for Language Learning Researchers. *Language Learning*, *65*(S1), 160–184. https://doi.org/10.1111/lang.12116

Hancock, G. R., Stapleton, L. M., & Arnold-Berkovits, I. (2009). The tenuousness of invariance tests within multisample covariance and mean structure models. In T. Teo & M. S. Khine

(Eds.), *Structural equation modeling in educational research: Concepts and applications* (pp. 137–174). Rotterdam, Netherlands: Sense Publishers.

Haynes, M., & Carr, T. H. (1990). Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 375–421). San Diego, CA: Academic Press.

Henning, G. H., Ghawaby, S. M., Saadalla, W. Z., El-Rifai, M. A., Hannallah, R. K., & Mattar, M. S. (1981). Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language. *TESOL Quarterly*, *15*(4), 457–466. https://doi.org/10.2307/3586486

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

Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), 1089–1115. https://doi.org/10.1017/S0142716420000193

Hui, B., Godfroid, A., & Elgort, I. (2022). *A construct validation study of time-sensitive word measures*. PsyArXiv. https://doi.org/10.31219/osf.io/dwjmn

IIBC (2022). *TOEIC Listening & Reading Test jukenshasuu no suii* [The change in the number of test-takers of TOEIC Listening & Reading Test]. https://21606703.fs1.hubspotusercontent-na1.net/hubfs/21606703/tlr_transition_2021.pdf

Irvine, P., Atai, P., & Oller, J. W., Jr. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning*, *24*(2), 245–252. https://doi.org/10.1111/j.1467-1770.1974.tb00506.x

JACET Basic Word Revision Committee (2016). *The new JACET list of 8000 basic words*. Kirihara Shoten.

Jeon, E. H. (2011). Contribution of morphological awareness to second-language reading comprehension. *The Modern Language Journal*, *95*(2), 217–235. https://doi.org/10.1111/j.1540-4781.2011.01179.x

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

Johnson, J. S. (1992). Critical period effects in second language acquisition: The effect of written versus auditory materials on the assessment of grammatical competence. *Language Learning*, *42*(2), 217–248. https://doi.org/10.1111/j.1467-1770.1992.tb00708.x

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99. https://doi.org/10.1016/0010-0285(89)90003-0

Khaldieh, S. A. (2001). The relationship between knowledge of Icraab, lexical knowledge, and reading comprehension of nonnative readers of Arabic. *The Modern Language Journal*, *85*(3), 416–431. https://doi.org/10.1111/0026-7902.00117

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

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

Kobayashi, Y. (2001). The learning of English at academic high schools in Japan: Students caught between exams and internationalisation. *The Language Learning Journal*, *23*(1), 67–72. https://doi.org/10.1080/09571730185200111

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

Komiya, C., & McLean, S. (2024). Which type of vocabulary-knowledge better correlates with reading-comprehension, meaning-recognition or meaning-recall?. *RELC Journal*, 1–13. https://doi.org/10.1177/00336882241246619

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lange, K., & Matthews, J. (2020). Exploring the relationships between L2 vocabulary knowledge, lexical segmentation, and L2 listening comprehension. *Studies in Second Language Learning and Teaching*, *10*(4), 723–749. https://doi.org/10.14746/ssllt.2020.10.4.4

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30. http://hdl.handle.net/10125/66648

Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, *31*(6), 475–483. https://doi.org/10.1016/j.appdev.2010.09.004

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

Masrai, A. (2020). Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review*, *11*(3), 423–447. https://doi.org/10.1515/applirev-2018-0106

Masrai, A. (2022). The relationship between two measures of L2 phonological vocabulary knowledge and L2 listening comprehension. *TESOL Journal*, *13*(1), 1–16. https://doi.org/10.1002/tesj.612

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, *52*, 1–13. https://doi.org/10.1016/j.system.2015.04.015

Matthews, J., Masrai, A., Lange, K., McLean, S., Alghamdi, E. A., Kim, Y. A., Shinhara, Y., & Tada, S. (2024). Exploring links between aural lexical knowledge and L2 listening in

Arabic and Japanese speakers: A close replication of Cheng, Matthews, Lange and McLean (2022). *TESOL Quarterly*, *58*(1), 63–90. https://doi.org/10.1002/tesq.3212

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741–760. https://doi.org/10.1177/1362168814567889

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389–411. https://doi.org/10.1177/0265532219898380

Meara, P. (2010). *EFL Vocabulary Tests* (2nd ed.). Swansea: Centre for Applied Language Studies, University College of Swansea.

Mecartty, F. H. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, *11*(2), 323–348.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners?. *The Canadian Modern Language Review*, *63*(1), 127–147. https://doi.org/10.3138/cmlr.63.1.127

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse & M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Bristol, Blue Ridge Summit: Multilingual Matters.

Mizumoto, A., & Shimamoto, T. (2008). A comparison of aural and written vocabulary size of Japanese EFL university learners. *Language Education & Technology*, *45*, 35–51. https://doi.org/10.24539/let.45.0_35

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

Muthén, L. K. (2012, August 6). *Residual covariance matrix not positive definite* [Online forum post]. Mplus Discussion. https://www.statmodel.com/discussion/messages/14/3223.html?1450187719

Muthén, L. K., & Muthén, B. O. (2017). Mplus: Statistical analysis with latent variables: User's guide (Version 8). Los Angeles, CA: Muthén & Muthén.

Nakanishi, N. (2019). *Phoneme counter* (Ver5.1). Kobe Gakuin University. https://noriko-nakanishi.com/phoneme/

Nassaji, H., & Geva, E. (1999). The contribution of phonological and orthographic processing skills to adult ESL reading: Evidence from native speakers of Farsi. *Applied Psycholinguistics*, *20*(2), 241–267. https://doi.org/10.1017/S0142716499002040

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

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, *12*(6), 445–449. https://doi.org/10.1111/1467-9280.00383

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Proctor, C. P., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology*, *97*(2), 246–256. https://doi.org/10.1037/0022-0663.97.2.246

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193

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

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York: Routledge.

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*(1), 43–55. https://doi.org/10.1080/01638538409544580

Robles-García, P., McLean, S., Stewart, J., Shin, J., & Sánchez-Gutiérrez, C. H. (2024). The development and initial validation of O-WSVLT, a meaning-recall Online L2 Spanish Vocabulary Levels Test. *Language Assessment Quarterly*, *21*(2), 181–205. https://doi.org/10.1080/15434303.2024.2311724

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2023). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological

vocabulary knowledge. *Studies in Second Language Acquisition*, 1–27. https://doi.org/10.1017/S027226312300044X

Samuel, A. G., & Frost, R. (2015). Lexical support for phonetic perception during nonnative spoken word recognition. *Psychonomic Bulletin & Review*, *22*, 1746–1752. https://doi.org/10.3758/s13423-015-0847-y

Satori, M. (2021). Effects of working memory on L2 linguistic knowledge and L2 listening comprehension. *Applied Psycholinguistics*, *42*(5), 1313–1340. https://doi.org/10.1017/S0142716421000345

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. S*tudies in Second Language Acquisition*, *31*(4), 577–607. https://doi.org/10.1017/S0272263109990039

Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2024). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*, *41*(1), 89–108. https://doi.org/10.1177/02655322231162853

Thompson, M. S., & Green, S. B. (2013). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 163–218). Charlotte, NC: Information Age Publishing.

Uchihara, T. (2023). How does the test modality of weekly quizzes influence learning the spoken forms of second language vocabulary?. *TESOL Quarterly*, *57*(2), 595–617. https://doi.org/10.1002/tesq.3176

Uchihara, T., & Harada, T. (2018). Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly*, *52*(3), 564–587. https://doi.org/10.1002/tesq.453

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

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*(3), 191–210. https://doi.org/10.1017/S0261444807004338

Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72*(1), 5–44. https://doi.org/10.1111/lang.12424

Wang, Y., & Treffers-Daller, J. (2017). Explaining listening comprehension among L2 learners of English: The contribution of general language proficiency, vocabulary knowledge and metacognitive awareness. *System*, *65*, 139–150. https://doi.org/10.1016/j.system.2016.12.013

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37. https://doi.org/10.1037/1082-989X.8.1.16

Wilson, I., Kaneko, E., Lyddon, P., Okamoto, K., & Ginsburg, J. (2011). Nonsense-syllable sound discrimination ability correlates with second language (L2) proficiency. In W. S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 2133–2136). Hong Kong: City University of Hong Kong.

Wong, S. W. L., Mok, P. P. K., Chung, K. K., Leung, V. W. H., Bishop, D. V. M., & Chow, B. W. (2017). Perception of native English reduced forms in Chinese learners: Its role in listening comprehension and its phonological correlates. *TESOL Quarterly*, *51*(1), 7–31. https://doi.org/10.1002/tesq.273

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696–725. https://doi.org/10.1177/1362168820913998