Full Research Article

The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis

Songshan Zhang [iD]
Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, China

Xian Zhang [iD]
Department of Linguistics, University of North Texas, USA

Abstract
This study set out to investigate the relationship between L2 vocabulary knowledge (VK) and second-language (L2) reading/listening comprehension. More than 100 individual studies were included in this meta-analysis, which generated 276 effect sizes from a sample of almost 21,000 learners. The current meta-analysis had several major findings. First, the overall correlation between VK and L2 reading comprehension was .57 ($p < .01$) and that between VK and L2 listening was .56 ($p < .01$). If the attenuation effect due to reliability of measures was taken into consideration, the 'true' correlation between VK and L2 reading/listening comprehension may likely fall within the range of .56–.67, accounting for 31%–45% variance in L2 comprehension. Second, all three mastery levels of form–meaning knowledge (meaning recognition, meaning recall, form recall) had moderate to high correlations with L2 reading and L2 listening. However, meaning recall knowledge had the strongest correlation with L2 reading comprehension and form recall had the strongest correlation with L2 listening comprehension, suggesting that different mastery levels of VK may contribute differently to L2 comprehension in different modalities. Third, both word association knowledge and morphological awareness (two aspects of vocabulary depth knowledge) had significant correlations with L2 reading and L2 listening. Fourth, the modality of VK measure was found to have a significant moderating effect on the correlation between VK and L2 text comprehension: orthographical VK measures had stronger correlations with L2 reading comprehension as compared to auditory VK measures. Auditory VK measures, however, were better predictors of L2 listening comprehension. Fifth, studies with a shorter script distance between L1 and L2 yielded higher correlations between VK and

Corresponding author:
Xian Zhang, Department of Linguistics, University of North Texas, Discovery Park Room B201, 3940 North Elm Street, Denton, TX 76201, USA.
Email: xian.zhang@unt.edu

L2 reading. Sixth, the number of items in vocabulary depth measures had a positive predictive power on the correlation between VK and L2 comprehension. Finally, correlations between VK and L2 reading/listening comprehension was found to be associated with two types of publication factors: year-of-publication and publication type. Implications of the findings were discussed.

Keywords
L2, listening, meta-analysis, reading, second language vocabulary, language acquisition, language teaching, vocabulary, vocabulary knowledge

I Introduction

Vocabulary knowledge (VK) plays a pivotal role in second-language (L2) reading comprehension (e.g. Bernhardt, 2011; Grabe, 2009; Nation, 2001) and L2 listening comprehension (Nation, 2001; Rost, 2013). A considerable level of vocabulary breadth, defined as 'the number of words for which the person knows at least some of the significant aspects of meaning knowledge' (Anderson & Freebody, 1981, p. 93), is necessary for learners to reach a certain lexical coverage rate needed for unassisted reading or listening comprehension (e.g. Adolphs & Schmitt, 2003; Hu & Nation, 2000; Nation, 2006). For reading comprehension, researchers proposed that a lexical coverage rate of 98%, which translates into a vocabulary size of 8,000–9,000 word families, would be reasonable for readers to adequately understand written texts (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Schmitt, Jiang, & Grabe, 2011). For listening comprehension, the vocabulary size needed to achieve the lexical coverage of 95%, 96%, and 98% is 3,000 word families, 5,000 word families, and 6,000–7,000 word families (Adolphs & Schmitt, 2003; Nation, 2006; van Zeeland & Schmitt, 2013) respectively.

The importance of VK can also be demonstrated by its correlation with L2 reading (e.g. Alderson, 2005; Jeon & Yamashita, 2014; Lervåg & Aukrust, 2010) and listening (Andringa et al., 2012; Matthews, 2018; Matthews & Cheng, 2015; McLean, Kramer, & Beglar, 2015; Vandergrift & Baker, 2015; Vandergrift & Baker, 2018). But it should also be noted that there is a large variation on the correlations between VK and L2 reading/listening comprehension across studies. Correlations between .08 and .95 were reported to exist between VK and L2 reading (e.g. Deacon, Wadewoolley, & Kirby, 2007; Droop & Verhoeven, 2003; van Gelderen et al., 2007). Similarly, VK and L2 listening comprehension are found to have a correlation range of .13–.85 (e.g. Guo, 2001; Proctor et al., 2005). These large variations could be attributed to many factors such as VK related factors (e.g. type of VK, modality of vocabulary measure), participant related factors (e.g. age, first language and second language), and publication related factors (year-of-publication, type of publication).

II VK related factors

1 Type of form–meaning knowledge

In practice, researchers operationalize vocabulary breadth knowledge as knowledge of form–meaning mapping (Schmitt, 2014). Laufer et al. (2004) and Laufer and Goldstein

(2004) distinguished four categories of form–meaning mapping knowledge: active recall, passive recall, active recognition, and passive recognition:

- Active recall is defined as the ability to produce the L2 form of a given meaning (e.g. word translation task).
- Passive recall involves the ability to provide the meaning for an L2 word (e.g. word definition task).
- Active recognition requires language users to recognize the L2 word form of a given meaning (e.g. word selection task based on a given meaning).
- Passive recognition requires language users to recognize the meaning of an L2 word (e.g. meaning selection task based on a given L2 word form).

These types of form–meaning knowledge were specified more explicitly by Schmitt (2010) with a different set of terms: form recall (active recall), form recognition (active recognition), meaning recall (passive recall), and meaning recognition (passive recognition). Schmitt's terms will be used in this article as these terms directly reveal the types of VK being denoted.

There is a growing interest in the field to investigate which types of meaning-form knowledge correlate better with L2 comprehension. There is some empirical evidence to suggest that different types of VK may correlate with L2 reading/listening differently (Cheng & Joshua, 2018; Laufer & Aviad-Levitzky, 2017). Laufer and Aviad-Levitzky (2017) found that meaning recognition tests were better predictors of L2 reading comprehension as compared to meaning recall tests. However, Cheng and Joshua (2018) suggested that form recall knowledge had a stronger correlation with L2 reading/listening than meaning recognition. The conflicting findings suggest that there is a need to further investigate this issue.

2   Type of vocabulary depth

It is widely acknowledged that VK is a complex, multidimensional construct (e.g. Daller, Milton, & Treffers-Daller, 2007; Milton, 2009; Milton & Fitzpatrick, 2014; Nation, 2001; Read, 2000; Schmitt, 2010, 2014). Scholars have formulated distinct conceptualizations of VK from different perspectives (e.g. Anderson & Freebody, 1981; Daller et al., 2007; Henriksen, 1999; Meara & Wolter, 2004; Nation, 2001; Read, 2000). Among various descriptive frameworks of VK, one of the most well-known frameworks is Anderson and Freebody's (1981) concept of vocabulary breadth and vocabulary depth (Qian, 1999, 2002; Qian & Schedl, 2004; Vermeer, 2001; Wesche & Paribakht, 1996). Vocabulary depth involves 'the quality or depth of understanding' (Anderson & Freebody, 1981, p. 93) and refers to how well a person knows the words (Qian, 2002). As vocabulary depth knowledge is a broad and loose construct that cannot be measured by one single test (Read, 2000; Schmitt, 2010, 2014), vocabulary specialists have formulated various conceptualizations about vocabulary depth knowledge (Nation, 2001; Qian, 1999; Read, 1993; Richards, 1976; for an overview, see also Schmitt, 2014). Among various aspects of vocabulary depth, morphological awareness and word association knowledge are two aspects of depth knowledge that have been frequently measured.

As its name indicates, morphological awareness is related to morphological knowledge of vocabulary items, which includes knowledge of the base unit, inflections, and affixes of vocabulary items. Associational knowledge is often operationalized as knowledge concerning whether a certain relationship exists between words. One good example is collocation, which refers to words that co-occur in texts (e.g. sound and sleep). Since morphological awareness and associational knowledge are measuring two different aspects of VK, we hypothesize that they could have different impacts on L2 reading and listening comprehension.

3   Modality of vocabulary measure

Phonological VK and orthographical VK are two important constructs of lexical knowledge. There is initial evidence indicating that learners' performances in phonological and orthographical vocabulary tests may differ (Milton & Hopkins, 2006; Milton & Riordan, 2006). It has been proposed that when investigating the relationship between VK and text comprehension, written vocabulary tests should be used in reading research while spoken vocabulary tests should be used in listening research (Milton, 2009; Read, 2000). However, it is often the case that spoken vocabulary tests, such as the Peabody Picture Vocabulary Test (PPVT), are used in L2 reading research. Likewise, written vocabulary tests, such as the Vocabulary Levels Test (VLT), are often used in L2 listening research. It remains unclear whether this mismatch in modality between VK measures and text comprehension tasks would moderate the relationship between VK and L2 text comprehension.

4   Context dependency of vocabulary measure

It remains debatable whether context-dependency of vocabulary measures would have an effect in predicting L2 text comprehension scores and in measuring L2 VK. Read (2000) and Read and Chapelle (2001) spoke in favor of contextualized vocabulary measures, though they have also suggested that the purpose of using a vocabulary test matters when designing and choosing the appropriate measures. This position was supported by Jeon and Yamashita (2014), who found that context-embedded vocabulary measures seemed to outperform context-independent vocabulary measures in predicting L2 reading comprehension. However, Qian (2008) did not find the effect of context on the predictive power of vocabulary measures on reading. He suggested that the discrete-point vocabulary measures and the contextualized vocabulary measures could have similar predictive value of L2 reading comprehension. This argument also found support from Henning (1991) and Qian and Schedl (2004). The two conflicting arguments in the literature motivated the current study to investigate the moderating effect of context dependency of vocabulary measure.

5   Number of test items in the vocabulary measure

The number of test items in the vocabulary measure is important in that it is closely related to the sampling rate of the vocabulary measure. Gyllstad, Vilkaitė, and Schmitt

(2015) explored how different sampling rates could affect the representativeness of each frequency band in the vocabulary size test. They found that 30 items per frequency band were adequate for most VK tests. However, no previous study has been carried out to investigate how sampling rate may moderate the correlation between VK and L2 comprehension.

III Participant related factors

1 Age

Age is a recurrent topic in second language acquisition. It is agreed that age is an important factor associated with language learning. While age is related to many aspects of language learning such as rate of acquisition and ultimate attainment (Birdsong, 2005; Ellis, 2008; Ortega, 2009), few studies have investigated the moderating effect of age on the relationship between VK and L2 comprehension. One notable exception is Jeon and Yamashita (2014), who revealed that the overall VK had a stronger correlation with L2 reading comprehension among adults/adolescents than children. Based on this finding, the current study intended to evaluate the moderating effect of age on the correlation between VK and L2 reading/listening comprehension.

2 L1–L2 distance

In a review of the effects of cross-linguistic factors on L2 reading comprehension, Koda and Reddy (2008) have argued that qualitatively different sub-skills are involved in reading typologically different languages. Language distance between first language (L1) and L2 (language distance) is related to L1 transfer (Oldin, 1989), a factor that is of great importance in the field of second language learning. Jeon and Yamashita (2014) hypothesized that L1–L2 distance might moderate the correlation between VK and L2 reading comprehension because of the cognate effect on L2 learning. Inspired by Jeon and Yamashita's hypothesis, L1–L2 distance was included in our moderator analysis to investigate whether L1–L2 distance could modulate the correlation between VK and L2 reading/listening comprehension. Following Jeon and Yamashita (2014) and Melby-Lervåg and Lervåg (2011), L1–L2 language distance was measured by two indices: language family and script distance (see below for a further discussion).

3 Context of L2 learning

One major question in second language acquisition is the effect of learning context on ultimate L2 success (e.g. Freed, 1995; Sanz, 2014). The context of L2 learning has been found to have an impact on L2 fluency development, phonological development, and morphosyntactic development (Sanz, 2014). In a recent study, Dóczi and Kormos (2016) investigated the development of L2 VK and lexical organization under the study-abroad context and the study-at-home-country context. They revealed different developmental patterns under the two learning conditions. Given the effect of L2 learning context on L2 VK development and the close relationship between VK and L2 comprehension, it is

reasonable to believe that the context of L2 learning can be a potential variable that moderates the correlation between VK and L2 comprehension.

IV Publication related factors

1 Year-of-publication

Year-of-publication may be related to the strength of correlation between VK and L2 comprehension. It is reasonable to believe that there are certain differences between studies conducted in recent years and studies in early years because advancement in VK measures could have benefited the studies in recent years. Moreover, studies in the past may be used as references for studies published at a later time so as to overcome limitations of earlier studies, which may lead to different findings between recent studies and earlier studies. As far as we know, no previous study has investigated whether year-of-publication is related to the correlation between VK and L2 reading/listening comprehension.

2 Publication type

Publication type has been claimed to be related to the effect sizes in meta-analysis because editors, reviewers, and researchers tend to favor studies with significant effect sizes (Cornell & Mulrow, 1999; Dickersin, 2005; Torgerson, 2006). Therefore, it is reasonable to believe that the correlation between VK and L2 comprehension could be affected by publication type, a hypothesis that has not been tested before.

3 Purpose

To our best knowledge, only one meta-analytic study (Jeon & Yamashita, 2014) was conducted to investigate the relationship between VK and L2 reading comprehension. Based on the meta-analysis of a sample of 31 studies, Jeon and Yamashita (2014) reported that the correlation between overall VK and L2 reading comprehension was .79. While the focus of Jeon and Yamashita (2014) was L2 reading comprehension, the impact of VK on L2 listening comprehension has not been studied.

Moderator variables in Jeon and Yamashita (2014) included two types of vocabulary measure characteristics (productive/receptive, embedded/discrete) that are relevant to the correlation between VK and L2 reading comprehension. While it is useful to distinguish between receptive and productive knowledge, the distinction between these two types of knowledge is not always straightforward (Schmitt, 2010, pp. 80–89; Waring, 1999). Read (2000) pointed out the problem of measuring receptive/productive knowledge by suggesting that receptive/productive knowledge has often been confused with comprehension/ use and different studies measured receptive/productive knowledge in different ways. One way to solve this measurement issue is to use the form–meaning recognition-recall model (Laufer & Goldstein, 2004), in which form–meaning entails types of knowledge and recognition-recall entails the types of task being performed (Schmitt, 2010).

How different types of VK (e.g. different types of form–meaning knowledge, morphological awareness) are related to L2 reading/listening comprehension is not yet

completely clear. The current study set out to investigate how VK is related to L2 reading and listening using meta-analysis, which offers a viable solution to examine questions that cannot be easily answered by a single empirical study. For this purpose, we intend to answer the following questions.

1. How are different types of VK related to L2 reading comprehension?
2. How are different types of VK related to L2 listening comprehension?
3. Which factors moderate the correlation between VK and L2 reading comprehension?
4. Which factors moderate the correlation between VK and L2 listening comprehension?

V Method

1 Identifying primary studies

When searching the literature for this meta-analysis, we followed the suggestions from Li, Shintanani, and Ellis (2012), In'nami and Koizumi (2012), and Oswald and Plonsky (2010). We adopted the literature searching practices reported in previous published meta-analyses pertaining to L2 comprehension. Five most frequently used online databases, including ERIC, LLBA, Web of Science, PsycINFO, and ProQuest Dissertations and Theses Full Texts, were searched to retrieve the primary studies. The following keywords and sets of their possible combinations were identified and used: vocabulary/word/lexical knowledge, vocabulary size/breadth knowledge, receptive/productive vocabulary knowledge, form recall/recognition, meaning recall/recognition, vocabulary depth knowledge, second language, foreign language, bilingual, reading/listening comprehension, text comprehension, and literacy development. Major reputable journals in the field of applied linguistics, reading and literacy development, and education were screened issue by issue online by the researchers. The references of relevant books, book chapters, review articles (narrative reviews versus meta-analyses) as well as individual studies were browsed to trace potential studies. In addition, search results from Google Scholar were also checked to ensure that the inclusiveness of the sample reached its saturation. These practices finally resulted in more than 5,600 potential studies that might include information relevant to this meta-analysis. The abstracts of these prospective studies were further carefully read by the researchers, and 244 out of them were identified that addressed the relationship between certain VK components and reading/listening comprehension. In the end, a total 116 studies (including 126 samples that would be treated as individual studies) that included at least one aspect of L2 VK as the predictor variable and L2 reading/listening comprehension as the criterion variable were included in this study. The literature retrieval procedure was given in Figure 1.

2 Coding scheme

Following the coding procedure for meta-analysis outlined by Lipsey and Wilson (2001, pp. 73–90), three parts of the coding scheme were distinguished, namely, Study Descriptors, Measurement Characteristics, and Effect Sizes. For ease of exposition, variables in each of the three categories were tabulated and elaborated in Table 1 .
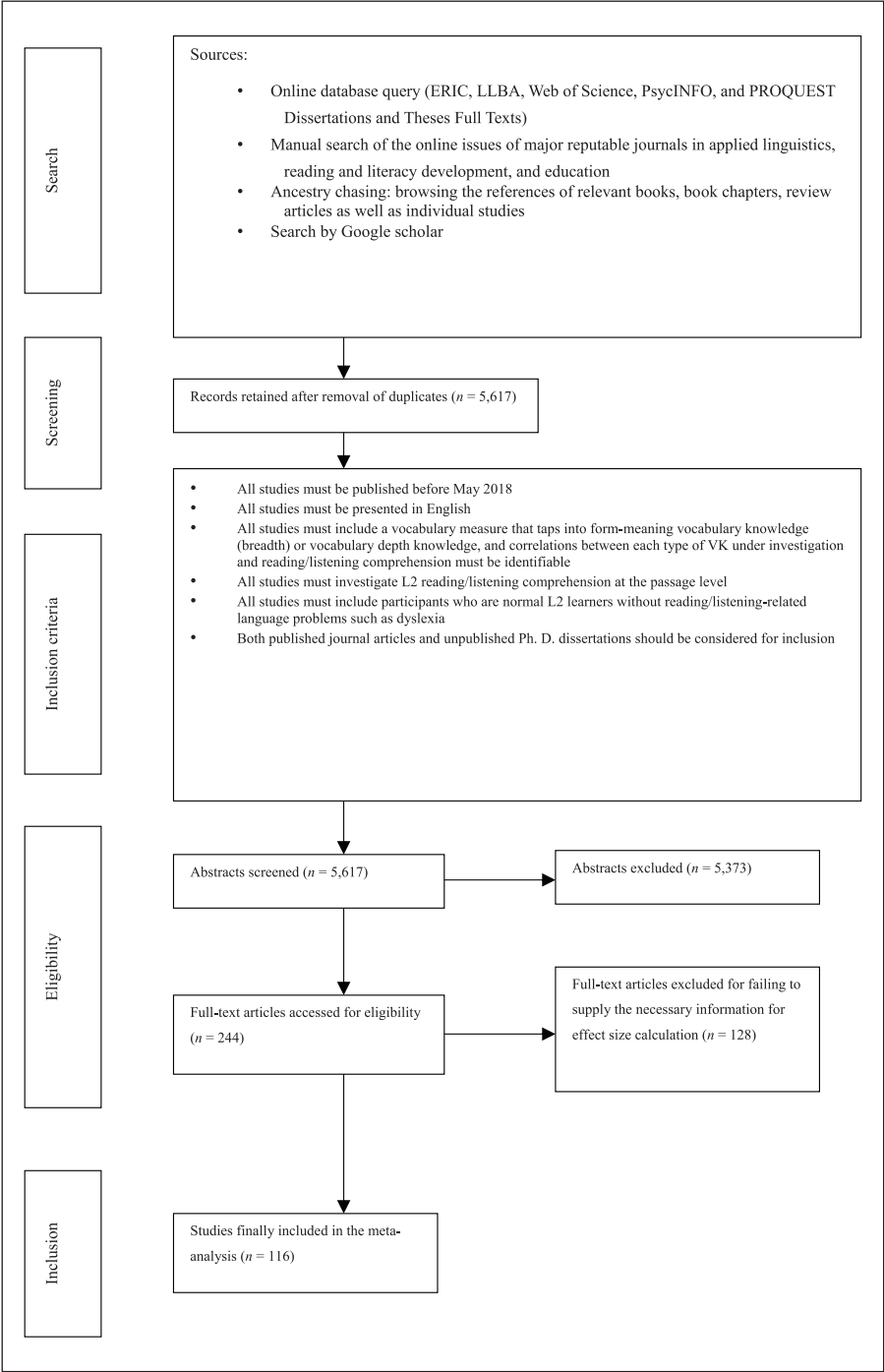
Sources:

- Online database query (ERIC, LLBA, Web of Science, PsycINFO, and PROQUEST Dissertations and Theses Full Texts)
- Manual search of the online issues of major reputable journals in applied linguistics, reading and literacy development, and education
- Ancestry chasing: browsing the references of relevant books, book chapters, review articles as well as individual studies
- Search by Google scholar

Search

Records retained after removal of duplicates (*n* = 5,617)

Screening

- All studies must be published before May 2018
- All studies must be presented in English
- All studies must include a vocabulary measure that taps into form-meaning vocabulary knowledge (breadth) or vocabulary depth knowledge, and correlations between each type of VK under investigation and reading/listening comprehension must be identifiable
- All studies must investigate L2 reading/listening comprehension at the passage level
- All studies must include participants who are normal L2 learners without reading/listening-related language problems such as dyslexia
- Both published journal articles and unpublished Ph. D. dissertations should be considered for inclusion

Inclusion criteria

Abstracts screened (*n* = 5,617)   Abstracts excluded (*n* = 5,373)

Eligibility

Full-text articles accessed for eligibility (*n* = 244)   Full-text articles excluded for failing to supply the necessary information for effect size calculation (*n* = 128)

Studies finally included in the meta-analysis (*n* = 116)

Inclusion

Figure 1. Flow chart for the literature retrieval procedures.

3   Moderator variables

Eleven variables in the Study Characteristics category and the Measurement Characteristics category were included in moderator analysis including year-of-publication, type of publication, age group of the participants, context of learning, language family of L1 and L2, script distance between L1 and L2, number of test items of the vocabulary measure, context-dependence of the vocabulary measure, modality of vocabulary measure, type of form–meaning knowledge, and type of vocabulary depth knowledge.

4   Coding reliability

All the studies were first coded three times by the first coder. Then 36 studies, about one third of the total sample, were randomly selected and brought to a second coder for coding. Following the procedure in previous meta-analyses (e.g. Boulton and Cobb, 2017; Plonsky, 2011), we calculated the percentage of agreement between the two coders. The

Table 1. Coding scheme for this meta-analysis.

Category	Variable	Definition and values
Study descriptors (*n* = 11)	Study ID	The order of the study, e.g. 1, 2, 3 . . .
	Author	The name(s) of the author(s)
	Year of publication (s)	The year the study was published, e.g. 2013
	Type of publication (c)	Journal article or thesis
	Time point (c)	The time point the vocabulary and reading/listening measures were administered, e.g. 1, 2, 3 . . .
	Age groups (c)	Primary school, middle school, or university
	Context of learning (c)	Foreign language context (if participants resided in a country where the target L2 was not the dominant language) or second language context (if participants resided in a country in which the target L2 was the dominant language)
	L1 of the participants (c)	Participants' first language, e.g. Spanish
	L2 of the participants (c)	Participants' second or foreign language, e.g. English
	Language family (of L1 and L2) (c)	SF (if L1 and L2 belonged to the same language family, e.g. Indo-European language) or DF (if L1 and L2 belonged to two different language families, e.g. one Indo-European language and the other non-Indo-European language)*
	Script distance (between L1 and L2) (c)	High Similarity (if both L1 and L2 were alphabetic languages or both non-alphabetic) or Low Similarity (if one of the languages was alphabetic and the other non-alphabetic)*

(Continued)

Table 1. (Continued)

Category	Variable	Definition and values
Measurement characteristics (*n* = 5)	Outcome measure (c)	Reading or listening
	Type of form–meaning knowledge (c)	Form recognition, form recall, meaning recognition, or meaning recall
	Type of vocabulary depth (c)	Morphological knowledge or word associational knowledge
	Number of items in depth measures (s)	A continuous variable recording the number of test items in each of the vocabulary measurement tools, e.g. 20, 25, 30 . . .
	Reliability of the outcome measure (s)	The reliability coefficients of the reading/listening measures as reported, e.g. .85, .93, .95 . . .
	Reliability of the vocabulary measure (s)	The reliability coefficients of different vocabulary measures as reported, e.g. .83, .91, .96 . . .
	Context-dependency of vocabulary measure (c)	Context-independent (if a vocabulary measure elicited test takers' response without providing any context) or context-dependent (if test takers produced responses by referring to contextual information)
	Modality of VK measure (c)	visual or auditory
Effect sizes (*n* = 3)	Correlation coefficient (s)	The correlation coefficient VK and reading/listening outcomes, e.g. r = .64
	Sample size (s)	Number of participants, e.g. 100, 120, 150 . . .
	Effect size direction	Positive or negative

Notes. * The classifications followed Jeon & Yamashita, 2014; Melby-Lervåg & Lervåg, 2011. (s) = continuous data type; (c) = categorical data type.

overall agreement between the two coders was 98%. All item-by-item agreement rates were above 94%. The agreement rates of some important variables such as age group, type of form–meaning knowledge, type of depth knowledge, context dependency, correlation coefficient, sample size, number of items in the VK measure, VK reliability, and context of learning were 99%, 94%, 97%, 99%, 98%, 99%, 99%, 98%, and 99%, respectively. The differences between the two coders were then resolved through discussion. Finally, the first coder checked the coding for multiple times to ensure coding accuracy.

5 Analysis

Statistical analysis was mainly carried out in the professional software package Comprehensive Meta-analysis (CMA, Borenstein et al., 2005). CMA offers two types of models to compute within-group effects (Q) and between-group effects ($Q_{between}$): fixed effects model and random effects model. Fixed effects model assumes that the true effect

is the same across studies and the heterogeneity in the observed effects is only due to sampling errors (Borenstein et al., 2011). Random effect model allows the true effect to vary across studies and the heterogeneity in the observed effects may be attributable to both between-study variations and sampling errors (Borenstein et al., 2011). Since the participants in the primary studies had diversified language and cultural backgrounds, the random effect model was chosen to compute the effect sizes. Following the common practice (Borenstein et al., 2011), all correlations were first converted to Fisher's z scores, which were used to compute the aggregated correlations and their confidence intervals. These results were then transformed back to correlations for presentation. Concerning the moderator analysis, all categorical variables were handled via between-group Q tests ($Q_{between}$) and continuous variables (e.g. year of publication) were analysed via meta-regression.

6   Multiple samples, multiple correlation coefficients

In the current meta-analysis, a majority of studies reported multiple correlations. To avoid sample size bias due to inclusion of multiple effect sizes from one study (Lipsey & Wilson, 2001), the current study adopted the common practice to ensure that each study contributed only one effect. In case that a study reported multiple correlations, these correlations would be pooled and the aggregated value would be used represent the effect size of the study. For example, Stæhr (2009) reported correlations between L2 listening and two types of VK. When computing the overall correlation between L2 listening and VK, the two correlation coefficients in Stæhr (2009) would be aggregated and only the pooled value would be used in the analysis. When there were multiple independent samples in a study, each independent sample would contribute one effect size as each independent sample would provide unique information and therefore could be treated as one individual study (Borenstein et al., 2011)

For moderator variable analyses that were carried out separately within the reading domain and the listening domain, studies that reported VK's correlations with both reading and listening would be used in their corresponding domains because these studies contributed only one effect size to the analysis in one domain. For example, Cheng and Joshua (2018) reported VK's correlation with L2 reading and L2 listening. The correlation with L2 reading would be used for the moderator analysis in the L2 reading domain and the correlation with L2 listening would be used in the L2 listening domain, which would not cause bias due to non-independent data points.

VI   Results

A total 126 independent studies (among the 116 primary studies, nine studies recruited multiple independent samples) were included in the current meta-analysis, generating 276 correlation coefficients from a sample of 20,969 participants, who had a diversified L1 background including Arabic, Chinese, Dutch, English, Farsi, Hebrew, Japanese, Korean, Spanish, Turkish, and Urdu etc. Among the 126 independent studies, thirty-one studies reported correlations of VK with both L2 reading and L2 listening. Seventy-nine studies reported correlations between VK and L2 reading comprehension only. Sixteen

Table 2. Reliability indices of the vocabulary knowledge (VK) and outcome measures.

Type of measures	K	Reported	Missing	Min	Max	Mean	SD	Median	IQR
VK	276	209	67	.54	.99	.85	.10	.86	.15
VK (i)	201	160	51	.54	.99	.84	.10	.86	.15
VK (ii)	75	49	26	.62	.98	.87	.09	.90	.13
L2 reading	110	82	28	.54	.97	.83	.09	.84	.10
L2 listening	47	29	18	.60	.97	.81	.09	.81	.13

Notes. VK (i) indicates VK measures administered in studies on L2 reading; VK (ii) indicates VK measures administered in studies on L2 listening.

studies reported correlations between VK and L2 listening comprehension only. In other words, there were a total 110 studies included in the L2 reading comprehension domain and 47 studies in the L2 listening domain.

Among the 276 correlations, the reliability indices of 209 correlations were reported (76%), with a median of .86. As for the outcome measures, about 75% reliability estimates of the L2 reading tests were reported (82/110). About 62% reliability estimates of the L2 listening tests were provided. The descriptive results were summarized in Table 2.

1 Outlier diagnosis

Outlier diagnosis was first performed to identify outliers. All correlation coefficients were entered into SPSS to perform outlier diagnosis. SPSS uses Tukey's hinges, which is based on the median, to calculate the interquartile range (IQR) of a dataset. Data points with 1.5 IQR or more from the upper Tukey's hinges (the 25 percentile above the median) or the lower hinge (the 25 percentile below the median) are identified as outliers. Among all L2 listening studies, no outlier was found. Among the reading studies, three studies were identified as outliers, including Droop and Verhoeven (2003), Hacking and Tschirner (2017), and Khaldieh (2001). Hoaglin and Iglewicz (1987) suggested that the criterion used by SPSS was too strict and proposed a more lenient criterion (2.2 as the multiplier instead of 1.5) for outlier detection. Using Hoaglin and Iglewicz's criterion, no study was diagnosed as an outlier. Therefore, all studies were included in the final analysis.

2 Overall effect

Table 3 summarizes the results for L2 reading and L2 listening. Both L2 reading and L2 listening comprehension had a moderate high correlation with VK: .57 ($Q = 1,322, p < .01$) and .56 ($Q = 560, p < .01$). For L2 reading, the real heterogeneity between studies accounted for 91.76% of the observed heterogeneity ($I^2 = 91.76, p < .01$). For L2 listening, a very high proportion of the observed heterogeneity ($I^2 = 91.79, p < .01$) was also made up by between studies differences. Since a high proportion of heterogeneity was attributional to between-study differences, subsequent moderator analysis was justified to identify the source of heterogeneity.

Table 3. Overall effects and publication bias.

	k	r	95% CI	Q	I²	Classic fail-safe N	Orwin's fail-safe N (trivial Fisher's z = .10)
L2 reading	110	.57**	.54–.61	1,322**	91.76	179,575	586
L2 listening	47	.56**	.50–.62	560**	91.79	31,205	247

Note. ** significant at .01 level.
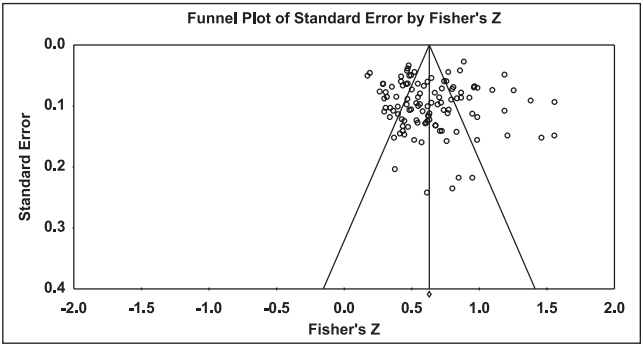
Figure 2. Funnel plot of studies involving second-language (L2) reading comprehension.

3 Publication bias

Publication bias was investigated using funnel plot, a commonly used method in meta-analysis for detecting publication bias. Figure 2 and Figure 3 display the funnel plots for studies in two domains. Each circle in the figures represents one individual study. Studies with large sample sizes and small standard errors would be displayed at the top of the graphs. Studies with small sample sizes and large standard errors would be located at the bottom of the graph. If a meta-analysis is able to capture all studies (indicating no publication bias), the funnel plot would show a symmetric pattern: studies are evenly dispersed on both sides of the overall effect (the mean, Borenstein et al. 2011), which is represented by the mid line in the triangle. If studies concentrate at the lower corners of the triangle, there is a potential publication bias (Borenstein et al. 2011). For both funnel plots, it can be noticed that a majority of studies were distributed at the upper parts of the triangles, suggesting that a majority of studies in the two domains had high precisions. In terms of symmetrical distribution, we did not find a clear sign of publication bias in the L2 listening domain: the funnel plot of Figure 3 is relatively symmetrically distributed. But for reading, the funnel plot (Figure 2) shows that the observed effect sizes were relatively unsymmetrical, suggesting that there could be a potential publication bias.

In addition to the funnel plot, more objective measures were also employed to identify potential publication bias. These measures included the trim-and-fill analysis (Duval & Tweedie, 2000), classic fail-safe N test, and Orwin's fail-safe N test (fail-safe N test is also known as the file-drawer analysis). The trim-and-fill analysis uses funnel plot to identify potential asymmetry of a summary effect. When asymmetry is found (more
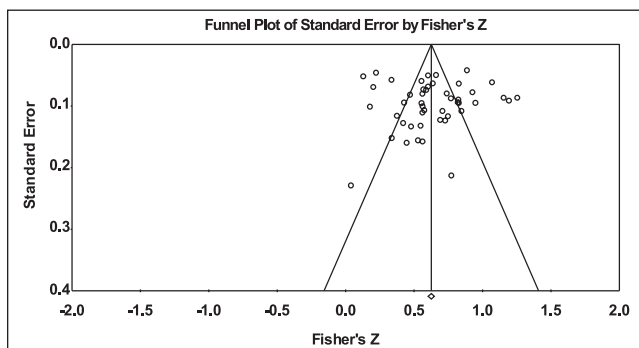
Figure 3. Funnel plot of studies involving second-language (L2) listening comprehension.

small studies on the right side of the mean than on the left), missing studies would be added to the analysis and the summary effect size would be re-computed. Rosenthal's Fail-safe *N* test calculates how many missing studies with a zero effect size are needed to be added to the meta-analysis in order to turn a significant effect into a non-significant one. If a large number of studies are required to nullify an effect, it is unlikely that the meta-analysis misses all these studies that can lead to a publication bias. Orwin's fail-safe *N* test works similarly to the classic Fail-safe *N* test. One main difference is that it is possible to set a critical value (e.g. smaller than .10) for missing studies in the Orwin's fail-safe *N* test whereas the critical value in the classic fail-safe *N* test is pre-set at nil.

Results of the trim-and-fill analysis showed that no missing study was needed to be added to the left of the mean in either domain. However, when imputed studies to the right of the mean were included, the overall effect sizes increased slightly from $r = .57$ (95% CI = [.54, .61]) to $r = .61$ (95% CI = [.58, .65]) for the L2 reading domain and from $r = .56$ (95% CI = [.50–.62]) to $r = .63$ (95% CI = .57, .68]) for the L2 listening domain. The improved correlations indicated that there could be a potential publication bias in favor of studies with smaller effect sizes. In other words, if the hypothetical effect sizes to the right of the means exist, we could observe stronger overall correlations between VK and L2 comprehension. In examining the results of the fail-safe *N* tests, we found that the numbers of missing studies to nullify the observed effect size in the two domains were both large (see Table 3). It is unlikely that the current meta-analysis missed such large numbers of studies. To sum up, a majority of studies in the current meta-analysis had large sample sizes and high precisions. The overall effect sizes were highly significant that the chance of finding a non-significant correlation was unlikely.

4 Moderator analysis

Table 4 summarizes the result of the moderator variable analysis. The moderator variables that had a significant impact on the correlation between VK and L2 reading included type of form–meaning knowledge, modality of vocabulary measure, script distance, and publication type. For L2 listening, variables that had a significant moderator effect included type of VK, and publication type.

Table 4. Moderator variable analysis in the second-language (L2) reading and L2 listening domain.

Moderators	L2 Reading				L2 listening			
	k	r	95% CI[3]	Q-test	k	r	95% CI[3]	Q-test
Type of FMK:								
Meaning recognition	14	.53	.49–.57	8.60*	22	.50	.41–.58	4.30
Meaning recall	21	.66	.58–.71		9	.58	.54–.62	
Form recall	66	.55	.48–.63		12	.63	.53–.72	
Recall vs. recognition:								
Recall	32	.61	.55–.66	5.09*	18	.61	.54–.69	4.36*
Recognition	73	.53	.49–.57		20	.49	.39–.58	
Type of vocabulary depth:								
Associational	15	.54	.43–.64	.05	3	.63	.23–.84	1.42
Morphological	18	.53	.44–.61		2	.32	−.13–.65	
Modality of VK test:								
Auditory	37	.49	.46–.54	9.02**	15	.60	.54–.65	3.08
Orthographical	72	.60	.54–.63		30	.52	.44–.59	
Context of VK items:								
Dependent	29	.59	.51–.66	.63	11	.54	.40–.65	.06
Independent	79	.55	.51–.59		36	.56	.49–.62	
Language families:								
SF	39	.54	.47–.60	.34	15	.58	.45–.68	.03
DF	50	.56	.51–.60		25	.56	.48–.64	
Script distance:								
Shorter distance	64	.58	.53–.63	3.85*	27	.60	.52–.67	3.04
Greater distance	28	.51	.45–.56		14	.49	.38–.59	
Research setting:								
Foreign language	42	.58	.52–.64	.20	22	.55	.45–.63	.24
Second language	68	.57	.53–.60		25	.58	.50–.64	
Age:								
University	39	.61	.54–.67	4.26	22	.53	.43–.62	1.29
Secondary	20	.59	.51–.67		7	.54	.40–.66	
Elementary	50	.53	.48–.58		17	.60	.43–.62	
Publication type:								
Dissertation	22	.50	.44–.56	6.13*	11	.39	.21–.56	5.84*
Journal	88	.59	.55–.63		36	.60	.55–.65	

Notes. FMK = form–meaning knowledge. VK = vocabulary knowledge. SF = same family. DF = different families. CI = confidence interval. ** significant at .01 level. * significant at .05 level. SF: if L1 and L2 belonged to the same language family, e.g. Indo-European language. DF: if L1 and L2 belonged to two different language families, e.g. one Indo-European language and the other non-Indo-European language.

a   Form–meaning knowledge. Since VK has multiple constructs (Nation, 2001), moderator variable analysis was performed to compare the effect of different types of VK on L2 reading and listening, including three mastery levels of form–meaning knowledge: meaning recognition, meaning recall, and form recall. All three types of form–meaning

knowledge had moderate to high correlations with L2 reading comprehension, with meaning recall having the strongest correlation ($r = .66$, $p < .01$), followed by form recall ($r = .55$, $p < .01$), and meaning recognition ($r = .53$, $p < .01$) . Regarding the question of which type of form–meaning knowledge is most critical for L2 reading comprehension, the result suggests that meaning recall may be the most important type of form–meaning knowledge for L2 reading as it had a higher correlation with L2 reading comprehension than meaning recognition ($Q_{between} = 9.44$, $p < .01$) and form recall ($Q_{between} = 5.88$, $p < .05$). Indeed, among various types of VK surveyed in the current meta-analysis (including vocabulary depth knowledge), meaning recall knowledge explained the largest proportion of variance in L2 reading comprehension (43.6%).

L2 listening comprehension was found to correlate with all three mastery levels of form–meaning knowledge, with form recall knowledge having the strongest correlation with L2 listening ($r = .63$, $p < .01$) and meaning recognition the weakest ($r = .50$, $p < .01$). Pairwise comparisons between the three mastery levels suggests that the effect of form recall on listening comprehension was stronger than meaning recognition ($Q_{between} = 4.85$, $p < .05$). Form recall had a slightly higher correlation as compared to meaning recall ($r = .58$) but the difference did not reach statistically significance ($Q_{between} = .52$, $p > .05$).

The patterns observed above suggest that recall knowledge could have a stronger correlation with L2 reading/listening comprehension as compared to recognition knowledge. To test this, an addition moderator analysis was carried out to compare recall and recognition knowledge, which confirmed the observation: the correlation between recall knowledge and L2 reading comprehension ($r = .61$) was statistically stronger than recognition knowledge ($r = .53$) ($Q_{between} = 5.09$, $p < .05$). Furthermore, the correlation between recall knowledge and L2 listening comprehension ($r = .61$) was also higher than recognition knowledge ($r = .49$) ($Q_{between} = 4.36$, $p < .05$).

b   Vocabulary depth. Moderator variable analysis was run to compare the moderator effect of two types of vocabulary depth: morphological awareness and associational knowledge. In the current context, morphological awareness mainly consisted of knowledge concerning derivational morphology, morpheme detection, and compound structure. Associational knowledge was mainly measured by collocational knowledge. L2 reading comprehension had a significant correlation with both associational knowledge ($r = .54$, $p < .01$) and morphological knowledge ($r = .53$, $p < .01$). No statistically significant difference was found between associational knowledge and morphological awareness in terms of their correlations with L2 reading ($Q_{between} = .00$, $p > .05$). For L2 listening, morphological awareness had a correlation of .32 ($p > .05$) with a CI of [−.13, .65]. Given the range of CI, the result concerning a significant correlation between morphological awareness and L2 listening comprehension is inconclusive. Associational knowledge, however, generated a much stronger correlation ($r = .63$, $p < .01$). Nonetheless, it should be noted that both pooled correlation coefficients were based on small sample sizes (n = 2, 3, Table 4) and more studies are needed to evaluate or confirm the result.

c   Modality of VK measure. For L2 reading, the modality of VK measure was a significant moderator. The aggregated correlation between orthographic VK measures and L2 reading comprehension ($r = .60$, $p < .01$) was greater than the correlation ($r = .49$,

$p < .01$) between auditory VK measures and reading comprehension ($Q_{between} = 9.02$, $p < .01$), suggesting that orthographic VK measures were better predictors of L2 reading comprehension. Within the L2 listening domain, auditory VK measures had a correlation of .60 with L2 listening comprehension. Orthographic VK measures, however, had a weaker correlation ($r = .52$, $p < .01$) with L2 listening comprehension. Although the difference (auditory vs. orthographic) did not reach the .05 significance level but the value was approaching the threshold ($Q_{between} = 3.08$, $p = .08$). Given that the number of studies in the L2 listening domain was relatively small, we believe that auditory VK measures can be a better predictor of L2 listening comprehension.

d Context dependency. Regarding context dependency of VK items in reading, the context-independent VK measures produced a slightly higher correlation ($r = .59$, $p < .01$) as compared to the context-dependent VK measures ($r =. 55$, $p < .01$). L2 listening, however, had a different pattern: the context-dependent VK measures generated a slightly higher correlation ($r = .56$, $p < .01$) as compared the context-independent VK measures ($r = .54$, $p < .01$). Nonetheless, both differences were not very large and did not reach the statistical significance level: reading ($Q_{between} = .63$, $p > .05$) and listening ($Q_{between} = .06$, $p > .05$).

e Language family. The impact of language family on the correlation turned out to be relatively small. For both L2 reading and listening, the correlations ranged between .54 and .58. The differences between SF and DF were small (about .02) with non-significant between-study Q for L2 reading ($Q_{between} = .34$, $p > .05$) and L2 listening ($Q_{between} = .03$, $p > .05$).

f Script distance. Script distance was found to have a significant moderator effect on reading. The aggregated correlation of the studies with a shorter script distance (e.g. both L1 and L2 were alphabetic languages) was .64 ($p < . 01$), higher than the r of .52 ($p < . 01$) of the studies with a greater script distance (one language is alphabetic and the other non-alphabetic) ($Q_{between} = 3.85$, $p < .05$). For L2 listening, a short script distance also seemed to benefit comprehension as the correlation was .60 ($p < . 01$) under the short distance condition. The long distance condition had a smaller correlation of .49 ($p < . 01$).

g Age. For L2 reading, the correlation among university students was .61 ($p < .01$), followed by secondary school students ($r = .59$, $p < .01$), and elementary school students ($r = .53$, $p < .01$). For L2 listening, the pattern reversed. The strongest correlation was found among elementary school students ($r = .60$, $p < .01$), followed by secondary school students ($r = .54$, $p < .01$), and university students ($r = .53$, $p < .01$).

h Context of studies. Regarding context of studies, no significant moderator effect was found on the correlation magnitude for either L2 reading comprehension ($Q_{between} = .20$, $p > .05$) or L2 listening comprehension ($Q_{between} = .24$, $p > .05$). For reading, the correlation in the foreign language context ($r = .58$, $p < .01$) was almost identical to the one in the second language context ($r = .57$, $p < .01$). For listening, the correlation in the
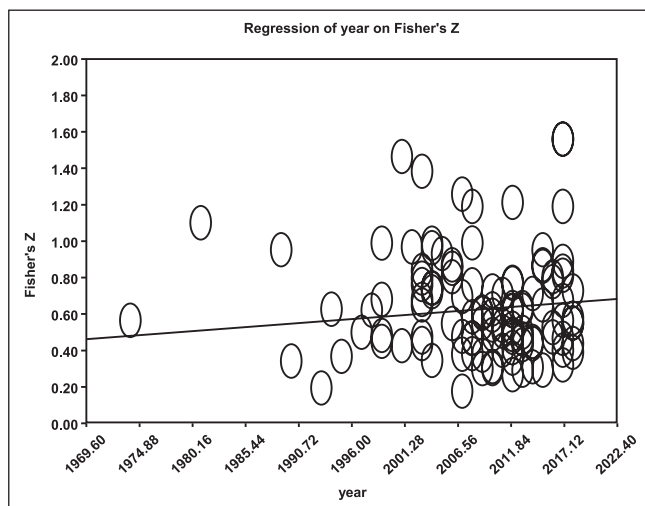
Figure 4. Effect of year-of-publication (second-language reading).

foreign language context ($r = .55, p < .01$) was slightly lower than the second language context ($r = .58, p < .01$).

i   Publication type. Publication type was found to be a significant moderator variable that affected the correlation between VK and L2 comprehension. For both L2 reading and listening comprehension, studies published in peer-reviewed journals tended to report higher correlations as compared to dissertation studies.

j   Year-of-publication. Since year-of-publication is a continuous variable, meta-regression was used to examine whether year-of-publication was related to the correlation strength between VK and L2 comprehension. Year-of-publication was found to be a significant positive moderator of the correlation between VK and L2 reading ($Q_{model} = 29.02$, B= .09, $p < .01$). Year-of-publication also affected the correlation magnitude between VK and L2 listening ($Q_{model} = 18.81$, B= .08, $p < .01$). Figure 4 and 5 depict the results of the meta-regression analysis.

5   Number of items in the vocabulary depth measures

L2 vocabulary breadth tests are often used to predict the number of vocabulary items known by a learner. Increasing number of items improves the accuracy of vocabulary tests as a measure of breadth knowledge (Gyllstad et al., 2015). On the other hand, vocabulary depth tests intend to measure how well vocabulary items are known. While some depth measures surveyed in the current meta-analysis included a relatively small number of items (some included eight items), the impact of a small item number on the correlation is not clear. In order to explore this issue, the moderator effect of number of
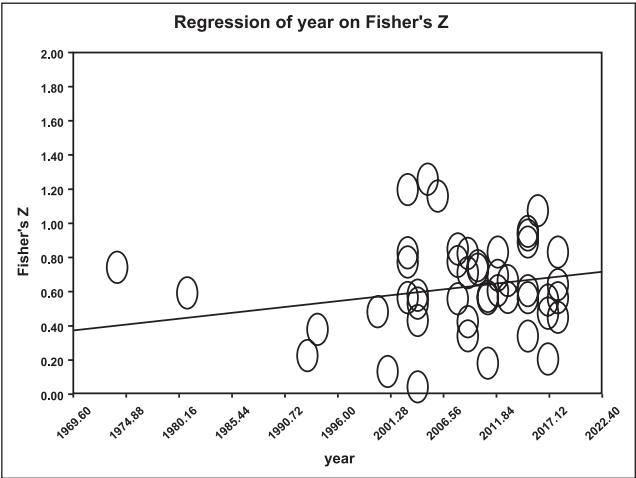
Figure 5. Effect of year-of-publication (second-language listening).
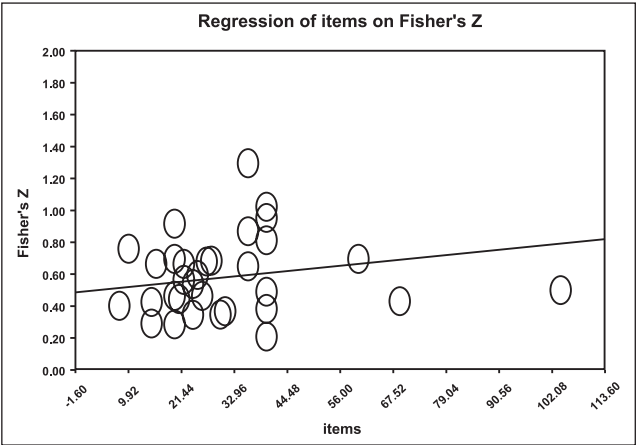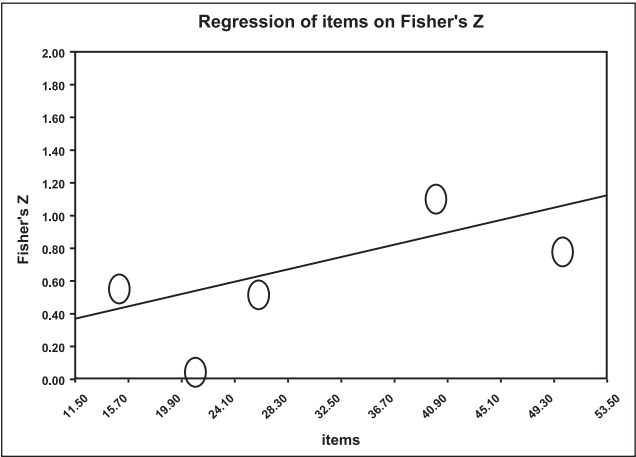

Figure 6. Effect of number of items (second-language reading).

items in the vocabulary depth measures (number of items) was investigated via meta-regression. Figure 6 and 7 give the scatter plots of the regressions. Number of items had a positive regression load on the correlation between VK and L2 reading comprehension (Q_{model} = 8.88, B = .003, p < .01). It also had a positive moderator effect on the correlation between VK and L2 listening comprehension (Q_{model} = 23.52, B = .02, p < .01).

Although including more items in depth tests could improve the predictive power of VK measures, it is not realistic to include hundreds of items in a vocabulary depth test. It is not clear how many vocabulary items are appropriate for a vocabulary depth test? No empirical evidence is available to answer this question. However, Gyllstad et al.

Figure 7. Effect of number of items (second-language listening).

Table 5. Moderator variable analysis on number of items.

Number of items	k	r	95% CI	$Q_{between}$
Group 1 ($<$ 30)	22	.47	.40–.54	5.19*
Group 2 ($\geqslant$ 30)	14	.62	.51–.71	

Note. * significant at .05 level.

(2015) suggested that a sample rate of 30 items in a frequency level (each frequency level contains 1,000 words) was appropriate to measure vocabulary breadth in the frequency level. Using this criterion, a moderator analysis was run on the number of items in vocabulary depth tests using 30 as the cutting point. We combined L2 reading and L2 listening in this analysis because there were only 5 studies in the L2 listening domain. Table 5 gives the result of the analysis. There was a significant group difference: the vocabulary depth tests with 30+ items produced a higher correlation ($r = .62, p < .01$) as compared to the other group ($r = .47, p < .01$) ($Q_{between} = 5.19, p < .05$) .

VII Discussion

Meta-analysis allowed the current study to investigate the role of VK on L2 comprehension based on a large number of participants (almost 21 thousand) with diversified language and cultural backgrounds, which, to a great extent, reduced the sampling bias. The aggregated correlation was .57 between VK and L2 reading and .56 between VK and L2 listening.

As mentioned earlier, a considerable proportion of studies failed to report the reliability indices for the VK and/or outcome measures. Regarding the reliability of instruments in the primary studies, this meta-analysis found the medians were .86, .84, and .81 for

VK measures, L2 reading tests, and L2 listening tests, respectively. These values came pretty close to those in Plonsky and Derrick (2016), who examined the reliability coefficients in second language research and revealed that the median reliability coefficients of VK measures, L2 reading tests, and L2 listening tests were .83, .86, and .77, respectively. If the reliability of instruments was to be taken into account to correct for the attenuation effect due to measurement errors, the adjusted correlations (corrected for attenuation) for reading and listening were both .67 (based on the double correction formula accounted for unreliability of both measures, see Muchinsky, 1996, for a description of the double correction formula and the single correction formula). It has been argued that unattenuated correlations are based on an ideal state of perfect reliable measures without measurement errors (Nunnally, 1994). Whereas the attenuated correlation may be an underestimate of the true effect due to measurement errors, the unattenuated correlation may be an overestimate of the true effect (Muchinsky, 1996). Thus the true effect of VK knowledge on L2 reading/listening comprehension may fall within the range between .56 and .67, accounting for 31%–45% variance in L2 reading/listening comprehension.

1 Different types of VK and their correlations with L2 reading and listening

There is little doubt that VK is important for both L2 reading and L2 listening. However, VK is a complex construct that includes many aspects of knowledge (Nation, 2001). While form–meaning knowledge is critical for L2 comprehension, we found that different mastery levels of form–meaning knowledge had a different impact on L2 reading and L2 listening. As discussed above, there was conflicting evidence concerning the role of different mastery levels of form–meaning knowledge in L2 reading. In fact, the ongoing debate centers on which mastery level of form–meaning knowledge is more critical for measuring VK for L2 reading. Nation's influential vocabulary size test (VST; Nation & Beglar, 2007) adopts the multiple-choice format that allows testees to pick an answer out of four meaning choices. According to the format of the test, the VST is a VK test that measures meaning recognition knowledge. A recent study by Laufer and Aviad-Levitzky (2017) argues that the meaning recognition test such as the VST is an adequate test to measure VK needed for reading comprehension. However, there is a concern (e.g. Stewart, 2014) that meaning recognition knowledge does not really reflect the needs of normal reading (e.g. reading novels, magazines, or newspapers) because in normal-reading situation meaning options are rarely offered for a reader to choose from. Indeed, in most of the natural situations, a reader has to actively recall the meaning(s) of vocabulary items during reading. Grabe (2009) argued that meaning recall knowledge is necessary for fluent reading because it requires automatic recall of vocabulary meanings so as to allocate more cognitive resources to comprehend reading texts. Therefore, meaning recall tasks such as L2 to L1 translation and word definition task may be the adequate measures of form–meaning knowledge for L2 reading (e.g. Stewart, 2014). The current study summarized the results of over 100 studies in L2 reading and found that meaning recall explained the largest proportion of variance in L2 reading comprehension,

suggesting that the mastery level of meaning recall may be most beneficial to L2 reading comprehension.

Compared to meaning recall, form recall knowledge had a weaker correlation with L2 reading comprehension. Form recall tasks require participants to produce the correct forms for vocabulary items (some of the typical form recall tasks include dictation and cloze). For fluent reading, form knowledge is necessary (Grabe, 2009; Perfetti & Hart, 2002; Stanovich, 2000) and a reader needs to first recognize a word so as to retrieve its meaning (Perfetti, 2007). Although it is desirable to recognize the correct form in order to be a fluent reader (Grabe, 2009), recalling the accurate forms of every word in a text may not always be necessary as context and partial knowledge (including partial knowledge of the form knowledge) can assist comprehension (Nation, 2001). It should be pointed out that form recall is the highest mastery level among the four types of form–meaning knowledge (Laufer & Goldstein, 2004). Laufer and Goldstein (2004) found that the vocabulary size based on form recall knowledge is smaller than the vocabulary size based on form recognition. Indeed, it is not uncommon that a L2 learner is able to recognize the correct form of an L2 vocabulary item but unable to produce the correct form.

Different from L2 reading comprehension, L2 listening comprehension seems to have a closer relationship with form recall knowledge. Among all types of VK surveyed in the current study, form recall knowledge had the strongest correlation with L2 listening comprehension. This result highlights the importance of form knowledge in L2 listening comprehension. Similar to L2 reading process within which the very first step involves recognizing the forms of words, L2 listening also starts from word recognition. However, visual word recognition (as in L2 reading) is different from spoken word recognition (as in L2 listening). While stimuli in reading involve words with clear boundaries (e.g. spaces and punctuations in English), stimuli in listening are a string of phonemes between which may or may not have clear boundaries. Listeners need to group phonemes to form words (the boundaries between words are not always acoustically marked) and match the words with the mental lexicon (Dahan & Magnuson, 2006). Without adequate form knowledge, spoken word segmentation (grouping phonemes into words) and recognition would become very difficult, not to mention comprehension.

2   Modality of VK measures

In theory, the modality of a VK measure (spoken/written) is sensitive to the modality of a comprehension task (Milton, 2009; Read, 2000). However, it is not uncommon in the field to measure auditory VK for L2 reading and orthographic VK for L2 listening. The impact of a mismatch in modality between a VK measure and an L2 comprehension task is not completely clear. The current study offers an answer to this question. The modality of VK measure was found to be related to the correlation between VK and L2 comprehension. L2 reading comprehension, a visual task, had a stronger correlation with orthographic VK measures as compared to auditory VK measures. L2 listening comprehension was found to have a higher correlation with auditory VK measures as compared to orthographic VK measures (approaching the significance threshold of .05). This result has a clear implication. When deciding which VK measure is to be used for L2 comprehension, the choice of a VK measure should take into account the modality of the comprehension task. Although

it is often assumed that if one modality (written or spoken) of VK is known the other modality is also likely known (Milton, 2009), this assumption does not necessarily hold as L2 learners' written and spoken VK can differ (Milton & Hopkins, 2006). Indeed, it is not impossible for an L2 learner to master a good amount of spoken vocabulary and a good L2 listening comprehension but know little about the written forms of vocabulary items (Milton, 2009). In this case, it is problematic to correlate auditory VK to L2 reading comprehension. In order to avoid the potential problem caused by mismatched modalities, we recommend researches in relevant areas to align the modality of a VK measure with the modality of a comprehension task.

3   Context dependency of vocabulary items

Regarding context dependency of vocabulary items in the vocabulary measures, we did not find a significant moderating effect of context dependency to the correlation between VK and L2 comprehension. This result is different from Jeon and Yamashita (2014), who found a significant moderator effect. This discrepancy could be due to a difference in the number of studies. The relevant analysis in Jeon and Yamashita (2014) involved a smaller number of studies (31 in total) as compared to 110 in the current meta-analysis. Taking a closer look at the studies that employed context dependent VK measures, we noticed that a majority of context-dependent VK measures embedded the target vocabulary items in single sentences, which provided only limited context for test takers. One reason to offer only limited context is because too much context may encourage inferencing and guessing, which is an undesirable artifact that can threaten the validity of a VK test. Little is known about the impact of limited context on the predictive power of VK measure on L2 comprehension. Our analysis provides initial evidence to suggest that embedding vocabulary items in limited context may not outperform the context-independent VK measures by much in terms of improving the correlation between VK and L2 comprehension.

4   Number of items in depth measures

Concerning the number of items in the vocabulary depth measures, the meta-regression analysis indicated that the number of items could moderate the correlation between VK and L2 reading/listening comprehension. Including more items in the vocabulary depth measures seemed to improve the explanatory power of VK on L2 reading/listening comprehension. Our analysis also suggested that vocabulary measures with 30+ items tended to generate a stronger correlation with L2 comprehension. Since the depth tests with 30+ items could explain a greater proportion of variance in L2 comprehension, we believe that 30+ may be an adequate number to measure vocabulary depth for L2 comprehension.

5   Age

Jeon and Yamashita (2014) investigated how age was related to the correlation between VK and L2 reading comprehension. How age may influence the relation between L2 listening comprehension has not been investigated before. The current study probed into this issue and found that the overall correlation between VK and L2 comprehension

among different age groups were in the range of .53–.61 for both L2 reading and L2 listening ($p < .01$), suggesting that VK played an important role for L2 comprehension among different age groups.

6   Language distance

Language distance between L1 and L2 (language distance) has been claimed to be related to L1 transfer (e.g. Odlin, 2003; Ringbom, 1987; Ringbom & Jarvis, 2009), a factor that is of great importance to L2 learning. Ringbom (1992) suggests that a shorter distance between L1 and L2 facilitates L2 comprehension (especially at the early stage of L2 learning) as lexical and grammatical knowledge shared by L1 and L2 can be employed to assist L2 comprehension. For example, Indo-European languages often have many cognates (a word in a language is regarded as a cognate to a word in another language if the two words have similar orthography and pronunciation). Cognates are found to be related to L2 processing (e.g. Lemhöfer & Dijkstra, 2004; Schwartz & Kroll, 2006) and L2 vocabulary acquisition (e.g. van Hell & de Groot, 2008). The current meta-analysis investigated whether language family can moderate the correlation between VK and L2 comprehension. If two languages belong to the same family (e.g. both L1 and L2 are Indo-European languages), they are believed to have a shorter language distance as compared to languages in two different language families (one Indo-European and one non-Indo-European). We did not find evidence to suggest that language family moderates the correlation between VK and L2 comprehension, which is consistent with Jeon and Yamashita (2014).

Among a number of language distance measures (e.g. language family, syntactic similarity), script distance may be the most relevant to the current study since the focus of the current study is vocabulary. When L1 and target L2 have a shorter script distance, word learning (defined as learning a target L2 form that shares certain similarities with an L1 form such as cognates) is beneficial to L2 learning and comprehension as learners may utilize L1 linguistic resources to compensate a lack of L2 linguistic resources that may hinder comprehension (Oldin, 1989). Indeed, the current study found that a short script distance could have a facilitating effect on the correlation between VK and L2 reading comprehension. The correlation between VK and L2 reading comprehension was larger among the participants whose L1 and L2 had a shorter script distance (e.g. both L1 and L2 were alphabetic languages) than the participants whose L1 and L2 had a longer script distance (e.g. L1 alphabetic and L2 ideographic). A shorter script distance between L1 and L2 allows learners to take advantage of L1 transfer, which may assist both L2 vocabulary acquisition and L2 comprehension (especially at the beginning stage of learning, Ringbom & Jarvis, 2009) and improves the explanatory power of VK on L2 comprehension (a higher correlation between VK and L2 comprehension among beginning learners could improve the overall correlation).

7   Publication relevant factors

Two publication relevant factors were analysed in the current meta-analysis: publication type and year-of-publication. Both factors were found to be related the correlation between VK and L2 reading/listening comprehension.

The studies published in journals reported an overall stronger correlation. The moderating effect was significant in both domains and the confidence intervals of the two types of publications barely overlapped in either domain. Dickersin (2005) suggested that studies in journals often reported larger effect sizes as compared to unpublished studies because studies accepted to referee journals usually had to go through a vigorous peer-review process and studies reported significant effects consistent with pre-established theories and expectations were more likely to be published in referee journals. The current meta-analysis found the same pattern for VK and L2 comprehension. Nevertheless, it shall be noted that such a pattern has not yet been fully tested with robust empirical evidence (e.g. a comparison of effect size between studies published in journals and studies rejected by journals), which, in our view, is an interesting topic to pursue.

Year-of-publication was positively related to the strength of correlation between VK and L2 reading/listening comprehension, suggesting that studies in recent years tended to report higher correlations between VK and L2 comprehension. The result is consistent with some previous meta-analysis that also found significant impacts of *Time* on their observed effect sizes (e.g. Kang & Han, 2015). One possible explanation to the *Time*'s moderating effect on the VK and L2 comprehension correlation may be that more comprehensive VK measures (e.g. Vocabulary Levels Test, Vocabulary Size Test, Word Association Test) have been developed, validated, and made available to L2 researchers in recent years. As more comprehensive VK measures are growing in number, researchers had more options to choose between VK measures, which could have improved the explanatory power of VK.

VIII   Implications and conclusions

The findings from the current study have a number of implications for language teaching and assessment. As meaning recall knowledge was found to have the highest correlation with L2 reading comprehension, teaching that focuses on meaning recall knowledge may best facilitate L2 reading comprehension. For example, if vocabulary cards are to be used to assist vocabulary learning along with meaning-focused incidental learning, the more effective way of using vocabulary cards is retrieval (e.g. translate L2 words) other than recognition (Nation, 2001). For L2 listening comprehension, however, the activities promoting form recall knowledge are recommended. For assessment purpose, meaning recall tests (e.g. a translation task) may be better VK measures in the L2 reading domain and form recall tests (e.g. partial dictation) may be better VK measures in the L2 listening domain. We also recommend that the modality of the VK test should match the modality of the L2 comprehension task (Milton, 2009) when investigating VK and L2 comprehension. Concerning the number of items in vocabulary depth tests, we found initial evidence to suggest that the number of 30+ may improve the predictive power of vocabulary depth.

Before we conclude the study, it is necessary to address a number of limitations. A review of the primary studies in the current meta-analysis showed that a majority of studies (more than 80%) involved English as the target L2, which prevented us from investigating the moderating impact of different L2s. We do believe that there is a great need for more studies targeting at languages other than English as the L2. In terms of

form–meaning knowledge, we also noticed that research investigating the mastery level of form recognition was largely missing from the literature, which restrained the current study from comparing all four mastery levels of form–meaning knowledge. Similarly, only two types of vocabulary depth were compared, mainly because they were the only types of vocabulary depth being measured. Although morphological awareness and associational knowledge are more popular in the literature, we want to emphasize that other types of vocabulary depth are no less important. In fact, it has been argued that every aspect of vocabulary depth is crucial (Nation, 2001; Schmitt, 2010; Webb, 2005). More studies are needed to evaluate the role of vocabulary depth in L2 comprehension, which will allow us to better understand the relationship between VK and L2 comprehension.

Declaration of conflicting interests

Also, if this study is part of a larger study or if you have used the same data in whole or in part in other papers, both already published or under review please state where the paper is published and describe clearly and in as much detail as you think necessary where the similarities and differences are and how the current manuscript makes a different and distinct contribution to the field.

ORCID iDs

Songshan Zhang ⃝iD https://orcid.org/0000-0003-2033-4561
Xian Zhang ⃝iD https://orcid.org/0000-0001-8472-5380

Supplemental material

Supplemental material for this article is available online.

References

Note. Studies that were also included in the meta-analysis are asterisked (*).

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*, 425–438.

Alderson, J.C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

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

*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*, 49–78.

Bernhardt, E.B. (2011). *Understanding advanced second-language reading*. London / New York: Routledge.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. In Kroll, J.F., & A.M.B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109–127). New York: Oxford University Press.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2005). *Comprehensive meta-analysis: Version 2*. Engelwood, NJ: Biostat.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2011). *Introduction to meta-analysis*. West Sussex: Wiley.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*, 348–393.

*Cheng, J., & Joshua, M. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*, 3–25.

Cornell, J., & Mulrow, C. (1999). Meta-analysis. In Herman, A., & M. Gideon (Eds.), *Research mythology in the social, behavioral, and life sciences* (pp. 285–323). London: Sage.

Dahan, D., & Magnuson, J.S. (2006). Spoken word recognition. In Traxler, M., & M.A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (pp. 249–283). San Diego, CA: Academic Press.

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

*Deacon, S.H., Wadewoolley, L., & Kirby, J. (2007). Crossover: The role of morphological awareness in French immersion children's reading. *Developmental Psychology*, *43*, 732–746.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origin and scope, and preventing harm. In Rothstein, H.R., Sutton, A.J., & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11–33). Chichester: Wiley.

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford: Oxford University Press.

*Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, *38*, 78–103.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

Ellis, R. (2008). *The study of second language acquisition*. 2nd edition. Oxford: Oxford University Press.

Freed, B.F. (1995). Language learning and study abroad. In Freed, B.F. (Ed.), *Second language acquisition in study abroad context*. Amsterdam / Philadelphia, PA: John Benjamins.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

*Guo, S. (2001). A multidimensional analysis of reading English as a second language by native speakers of Chinese. Unpublished PhD dissertation, University of Iowa, Iowa City, USA.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *IJAL – International Journal of Applied Linguistics*, *66*, 278–306.

*Hacking, J.F., & Tschirner, E. (2017). The contribution of vocabulary knowledge to reading proficiency: The case of college Russian. *Foreign Language Annals*, *50*, 500–518.

Henning, G. (1991). *A study of the effects of contextualization and familiarization on responses to TOEFL vocabulary test items*. Princeton, NJ: Education Testing Service.

Henriksen, B. (1999). Three dimensions of vocabulary knowledge. *Studies in Second Language Acquisition*, *21*, 303–317.

Hoaglin, D.C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*, 1147–1149.

Hu, M., & Nation, I.S.P. (2000). Unknown vocabulary density and reading comprehension. *Reading in A Foreign Language*, *23*, 403–430.

In'nami, Y., & Koizumi, R. (2012). Database selection guidelines for Meta-analysis in Applied Linguistics. *TESOL Quarterly*, *44*, 169–184.

Jeon, E.H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*, 160–212.

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A Meta-analysis. *The Modern Language Journal*, *99*, 1–18.

*Khaldieh, S.A. (2001). The relationship between knowledge of *I'raab*, lexical knowledge, and reading comprehension of nonnative readers of Arabic. *The Modern Language Journal*, *85*, 416–431.

Koda, K., & Reddy, P. (2008). Cross-linguistic transfer in second language reading. *Language Teaching*, *41*, 497–508.

*Laufer, B., & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, *101*, 729–741.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436.

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*, 15–30.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*, 202–226.

Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, *32*, 533–550.

*Lervåg, A., & Aukrust, V.G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry*, *51*, 612–620.

Li, S., Shintanani, N., & Ellis, R. (2012). Doing meta-analysis in SLA: Practices, choices, and standards. *Contemporary Foreign Language Studies*, *384*, 1–17.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

*Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System*, *72*, 23–36.

*Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, *52*, 1–13.

*McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*, 741–760.

Meara, P., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. In Albrechtsen, D., Haastrup, K., & B. Henriksen (Eds.), *Angles on the English-speaking World: Volume 4* (pp. 85–96). Copenhagen: Museum Tusculanum Press.

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, *34*, 114–135.

Milton, J. (2009). *Measuring second language vocabulary acquisition: Volume 45*. Bristol: Multilingual Matters.

Milton, J., & Fitzpatrick, T. (2014). Introduction: Deconstructing vocabulary knowledge. In Milton, J., & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge*. Basingstoke: Palgrave Macmillan.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *Canadian Modern Language Review*, *63*, 127–147.

Milton, J., & Riordan, O. (2006). Level and script effects in the phonological and orthographic vocabulary size of Arabic and Farsi speakers. In Davidson, P., Coombe, C., Lloyd, D., & D. Palfreyman (Eds.), *Teaching and learning vocabulary in another language* (pp. 122–133). United Arab Emirates: TESOL Arabia.

Muchinsky, P.M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63–75.

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

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–81.

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

Nunnally, J.C. (1994). *Psychometric theory*. 3rd edition. New York: McGraw-Hill.

Odlin, T. (2003). Crosslinguistic influence. In Doughty, C., & M. Long (Eds.), *The handbook of second language acquisition* (pp. 436–486). Oxford: Blackwell.

Oldin, T. (1989). *Language transfer*. Cambridge: Cambridge University Press.

Ortega, L. (2009). *Understanding second language acquisition*. Oxford: Oxford University Press.

Oswald, F.L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110.

Perfetti, C.A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*, 357–383.

Perfetti, C.A., & Hart, L. (2002). The lexical quality hypothesis. In Verhoeven, L., Elbro, C., & P. Reitsma (Eds.), *Precursors of functional literacy*. Philadelphia, PA: John Benjamins.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning*, *61*, 993–1038.

Plonsky, L., & Derrick, D.J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*, 538–553.

*Proctor, C.P., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-speaking children reading in English. *Journal of Educational Psychology*, *97*, 246–256.

*Qian, D.D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*, 282–308.

*Qian, D.D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*, 513–536.

*Qian, D.D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly*, *5*, 1–19.

*Qian, D.D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*, 28–52.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355–371.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J., & Chapelle, C.A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*, 1–32.

Richards, J.C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*, 77–89.

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Clevedon: Multilingual Matters.

Ringbom, H. (1992). On L1 transfer in L2 comprehension and L2 production. *Language Learning*, *42*, 85–112.

Ringbom, H., & Jarvis, S. (2009). The importance of cross-linguistic similarity in foreign language learning. In Doughty, C., & M. Long (Eds.), *The handbook of language teaching* (pp. 106–118). Oxford, UK: Wiley-Blackwell.

Rost, M. (2013). *Teaching and researching listening* (2nd ed.). London and New York: Routledge.

Sanz, C. (2014). Contributions of study abroad research to our understanding of SLA processes and outcomes: The SALA Project, an appraisal In Pérez-Vidal, C. (Ed.), *Language acquisition in study abroad and formal instruction contexts*. Amsterdam / Philadelphia, PA: John Benjamins.

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

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*, 913–951.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43.

Schwartz, A.I., & Kroll, J.F. (2006). Bilingual lexical activation in sentence context. *Journal of Memory & Language*, *55*, 197–212.

*Stæhr, L.S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*, 577–607.

Stanovich, K.E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford Press.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*, 271–282.

Torgerson, C.J. (2006). Publication bias: The Achilles' Heel of systematic reviews? *British Journal of Educational Studies*, *54*, 89–102.

*van Gelderen, A., Schoonen, R., Stoel, R.D., de Glopper, K., & Hulstijn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology*, *99*, 477–491.

van Hell, J., & de Groot, A. (2008). Sentence context modulates visual word recognition and translation in bilinguals. *Acta Psychologica*, *128*, 431–451.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479.

*Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*, 390–416.

*Vandergrift, L., & Baker, S.C. (2018). Learner variables important for success in L2 listening comprehension in French immersion classrooms. *Canadian Modern Language Review*, *74*, 79–100.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217–234.

Waring, R. (1999). Tasks for assessing second language receptive and productive vocabulary. Unpublished PhD dissertation, University of Wales Swansea, Swansea, UK.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on Word Knowledge. *Studies in Second Language Acquisition*, *27*, 33–52.

Wesche, M., & Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*, 13–40.