Accepted for publication in *The Language Learning Journal* (Routledge)

The role of spoken vocabulary knowledge in second language speaking proficiency

Takumi Uchihara and Jon Clenton

Abstract

Research has started revealing the important role of vocabulary knowledge in second language (L2) speaking proficiency. However, the majority of earlier studies tended to disregard the congruence in test format between assessing vocabulary knowledge and speaking skills with the former predominantly measured in written format. The current study therefore measured vocabulary knowledge in spoken format to university students speaking English as an L2, and investigated whether spoken vocabulary knowledge predicts speaking proficiency. Forty-six university learners completed written and spoken forms of productive vocabulary test (Lex30; Meara and Fitzpatrick 2000) as well as a story narrative task. Elicited speech samples were rated in terms of four aspects of L2 speaking proficiency (fluency, vocabulary, pronunciation, grammar), and the rating scores were compared with productive vocabulary scores. Results showed a significant correlation between the spoken and written vocabulary scores with a closer examination of the data indicating a gap between the two forms. Results of the vocabulary-speaking link indicated that spoken vocabulary knowledge was associated with all but one of the L2 speech ratings, while written vocabulary knowledge was not related to any of the rating scores. The current study provided methodological and practical implications with respect to the central role of modality in vocabulary testing.

Keywords: vocabulary knowledge, vocabulary testing, productive vocabulary, speaking proficiency, modality

Introduction

It is widely accepted that vocabulary knowledge is closely related to second language (L2) proficiency and development (Meara 1996; Qian and Lin 2020). The central role for vocabulary has been empirically supported by a growing number of studies showing the important relationship between vocabulary knowledge and L2 comprehension skills such as reading (e.g., Qian 1999, 2002; Laufer and Aviad-Levitzky 2017; McLean et al. 2020) and listening (e.g., Noreillie et al. 2018; Stæhr 2009; Vafaee and Suzuki 2020; Vandergrift and Baker 2015). Two meta-analyses by Jeon and Yamashita (2014) and Zhang and Zhang (2020) show medium-to-high correlations between vocabulary knowledge and comprehension skills with an average effect size of .57 and .79 for reading, and .56 for listening. Such evidence, along with research reporting predictive validity of vocabulary size for L2 general proficiency (e.g., Zareva et al. 2005) has in turn informed L2 instruction and assessment, offering useful tools for teachers to gauge general L2 proficiency more rapidly than administering large-scale standardized L2 proficiency exams such as TOEFL and IELTS (Milton 2010). In particular, vocabulary measures can serve as practical tools for diagnostic and placement purposes (Meara and Miralpeix 2016; Webb et al. 2017). However, the vast majority of studies focusing on the link between L2 vocabulary and proficiency have focused on L2 comprehension skills (Zhang and Zhang 2020). Limited attention has been paid to the predictive role of vocabulary knowledge for productive language skills (Miralpeix and Muñoz 2018); for example, Baba (2009) has looked at writing while Noreillie et al. (2020) have focused on speaking. More research on the link between speaking proficiency and vocabulary knowledge would be helpful for practical reasons as assessment of speaking proficiency is typically labour-intensive.

Emerging research has begun to reveal an important relationship between vocabulary knowledge and L2 oral proficiency (e.g., Clenton et al. 2021; Noreillie et al. 2020; Uchihara 2021). However, virtually all available studies have assessed vocabulary knowledge in written form (e.g., de Jong et al. 2012; Uchihara and Clenton 2020). The selection of written vocabulary tests is of course reasonable when the focus is on written L2 skills such as reading and writing (Baba 2009; Harrington and Carey 2009; Qian, 1999 2020) with congruency maintained between test and skill . In L2 speaking research, however, there has been minimal attention to congruency in test format, aside from two studies (Alhazmi and Milton 2016; Milton et al. 2010). This is problematic because written and spoken word knowledge, while related, are independent constructs (Cheng and Matthews 2018) and the developmental trajectory for written vocabulary is different from that for oral vocabulary (Zaytseva et al. 2021). Test incongruency may well lead to inaccurate estimations of learners' L2 proficiency (Jelani and Boers 2018); for example, if written vocabulary scores are used to predict oral proficiency, there may be underestimation in the case of learners with a large spoken vocabulary but small written vocabulary, and overestimation in the converse case. The role of modality is increasingly considered an important methodological component in research focusing on the link between L2 vocabulary and listening (e.g., McLean et al. 2015; Zhang and Zhang 2020), but has largely been neglected in research focusing on L2 vocabulary and L2 speaking.

In response to this research gap, the current study is primarily designed to explore the role of vocabulary knowledge, measured in both written and spoken format, in various aspects of L2 speaking proficiency. The secondary aim of this study is to investigate further the relationship the two forms of word knowledge (written and spoken). Since earlier studies comparing spoken and written modes focus on *receptive* vocabulary knowledge (Chen and Matthews 2018; Milton and Hopkins 2006; Uchihara and Harada 2018), the current study focuses on the different modes of *productive* vocabulary knowledge, which might provide additional insights into the role of modality in L2 vocabulary acquisition.

Testing vocabulary knowledge

Although researchers and practitioners agree that word knowledge is a multifaced construct (Fitzpatrick and Clenton 2017; Nation 2013; Yanagisawa and Webb 2020), it is the form-meaning connection of L2 words that has so far received the greatest amount of attention in vocabulary research. The majority of available tests purport to assess the number of words whose meanings (or forms) learners can recognise or comprehend (i.e., receptive vocabulary knowledge). Receptive knowledge is measured via a wide range of recognition tasks, including lexical decision (Harrington and Carey 2009), word-matching (e.g., Webb et al. 2017), and multiple-choice (Beglar 2010), as an essential linguistic component ofr L2 reading and listening skills (Jeon and Yamashita 2014; Zhang and Zhang 2020). In addition to receptive knowledge, learners may also need to acquire the ability to recall the forms (or meanings) of L2 words (i.e., productive vocabulary knowledge) so that they can produce L2 words in writing and speaking. Recall of forms is particularly relevant to the development of written and spoken production skills, since the spelling and pronunciation of words need to be sufficiently accurate and intelligible to interlocutors (Uchihara, 2022). Productive knowledge is often measured via free production tasks (e.g., essay writing; Laufer and Nation, 1995) or controlled production tasks, such as translation (Koizumi and In'nami 2013), gap-filling (Laufer and Nation, 1999), and word association (Meara and Fitzpatrick 2000) elicitation tasks.

A further complication in eliciting vocabulary knowledge relates to another important aspect—modality (spoken or written)—further dividing receptive and productive aspects of word knowledge into four components (spoken receptive, spoken productive, written receptive, and written productive). Research comparing written and spoken receptive word knowledge (Milton and Hopkins 2006; Uchihara and Harada 2018) suggests that vocabulary size in spoken form cannot be estimated directly from the results of written vocabulary measures. Milton and Hopkins (2006) measured knowledge of written and spoken forms of L2 words receptively with Arabic and Greek learners of English using a yes/no vocabulary recognition test. A strong correlation was observed (r = .688) between the written and spoken versions of the vocabulary test but the score for the spoken version of the test was significantly lower than that for the written version of the same test. These findings confirm a discrepancy between the two modes of vocabulary knowledge, in line with the study by Uchihara and Harada (2018) finding that Japanese learners of English showed smaller

vocabulary size in spoken form than in written form. Similarly, Cheng and Matthews (2018) also support the important role of modality in vocabulary testing. The researchers assessed written receptive (word-matching), written productive (gap-filling), and spoken receptive (dictation) vocabulary knowledge of 250 Chinese students studying English as a foreign language. Their factor analysis demonstrated that written and spoken knowledge are related but independent constructs. These findings accordingly suggest that the spoken-written modality, in addition to the receptive-productive dimension, needs further consideration in vocabulary testing (Fitzpatrick and Clenton 2010).

L2 Speaking and vocabulary

L2 speech research has consistently suggested the importanceof vocabulary as a strong predictor of L2 speaking proficiency even after other factors related to linguistic knowledge and processing skills are accounted for (de Jong et al. 2012; Iwashita et al. 2008; Saito et al. 2016). This line of research can be broadly divided into two approaches. The first focuses on a wide range of properties of L2 words appearing in speakers' production (e.g., lexical sophistication) as an indication of word knowledge. For example, the use of lower frequency words is regarded as evidence supporting advanced lexical knowledge (Kyle 2020). The second approach measures vocabulary and speaking separately and examines whether vocabulary test scores predict speaking proficiency. Although in the first approach the vocabulary and speaking measures are dependent in so far as they share the same source of elicited speech data, in the second approach the data for each construct are assessed independently and are therefore less subject to the issue of circularity (see Uchihara 2021 for a review). The latter will be reviewed in this section as it is more closely relevant to the goal of the current study.

From a theoretical standpoint, vocabulary knowledge is considered crucial for speaking proficiency. According to first language (L1) and L2 speech production models (Kormos 2006; Levelt, 1989), generating pre-verbal messages (*what* to say) triggers subsequent formulation of speech utterances (*how* to say). At the formulation stage, words that are conceptually appropriate to convey the messages are selected and retrieved. Lexical selection triggers syntactic building and phonological details are then specified before the final speech product is articulated as overt speech. Failure to access and retrieve L2 words could disturb the formulation stage with the entire processing more cognitively demanding and less efficient (Skehan 2009). Articulated L2 speech consequently might sound disfluent and linguistically inaccurate (Koizumi and In'nami 2013). In this sense, speech production is regarded as lexically driven (Kormos 2006; Levelt, 1989) and a well developed lexicon is a prerequisite condition for successful L2 speech production (Skehan 2009).

It is also important to note that the construct of speaking proficiency is complex (Iwashita et al. 2008; Saito et al. 2016) to the extent that some aspects of L2 speech might be more closely related to vocabulary knowledge than other oral features. For example, learners' lexical knowledge might be more directly related to lexical, grammatical, and temporal features compared to pragmatic or

discourse features. Larger vocabulary knowledge might enable correct word choice, and efficient processing as a result of rapid retrieval of L2 words might allow for a rapid and accurate allocation of attentional resources to sentence building and speech articulation (Kormos 2006; Skehan 2009). However, the ability to organize sentences in a cohesive and coherent manner (discourse competence) or evaluate the politeness conveyed through produced language (pragmatic competence) might be only remotely related to vocabulary knowledge. Accordingly, it is methodologically important to select specific aspects of L2 speaking proficiency that are theoretically relevant to learners' vocabulary knowledge.

Table 1. Summary of ~~some~~ earlier studies reporting correlation coefficients between vocabulary knowledge and L2 speaking measures

Study	Vocabulary test	Speaking measure	Result (*r* or *rho*)
Clenton et al. (2021)	checklist	fluency	|.02 to .24|
	WA	fluency	|.02 to .39|
de Jong and Mora (2017)	checklist	fluency	|.227 to .311|
Hilton (2008)	Dialang	fluency	|.390 to .668|
Mariano and Santiago (2020)	Dialang	fluency	.30 to .33
		pronunciation (rating)	.35
		pronunciation (acoustic)	|.01 to .19|
Miralpeix and Muñoz (2018)	checklist	fluency	.485
		pronunciation	.311
		lexico-grammar	.413
Noreillie et al. (2020)	MC	lexical richness	.19, .32
	gap-filling	lexical richness	.05, .27
Uchihara and Clenton (2020)	checklist	lexical richness	.173 to .552
Uchihara et al. (2021)	WA	fluency	|.03 to .48|
		lexical richness	|.10 to .47|
		pronunciation	|.23, .35|
Uchihara and Saito (2019)	WA	fluency	.342
		pronunciation	.034 to .271

Note. MC = multiple choice; WA = word association (i.e., Lex30); checklist = yes/no checklist format such as X_Lex and Y_Lex. Combined measure = speaking score based on fluency, vocabulary, grammar, and pronunciation.

This assumption that vocabulary knowledge plays an integral part in speech production has been supported by a growing number of empirical studies measuring various aspects of L2 speech considered relevant to learners' vocabulary knowledge (see Table 1 for a summary of earlier studies). Studies of oral fluency have supported this theoretical account of the key role of vocabulary in L2 speaking proficiency (Clenton et al. 2021; de Jong and Mora 2019; Hilton 2008; Koizumi and In'nami

2013; Miralpeix and Muñoz 2018; Uchihara et al. 2021; Uchihara and Saito 2019). Meanwhile, Hilton (2008) found medium-to-large correlations between DIALANG vocabulary test scores and a range of utterance fluency measures with the largest effect size observed for mean length of run (*rho* = .668). de Jong and Mora (2019) assessed receptive vocabulary knowledge through a checklist task (X/Y_Lex; Meara and Miralpeix 2016) and measured various fluency aspects with 51 upper-intermediate adult L1 Spanish learners of English. A significant correlation was found between vocabulary test scores and mean syllable duration (r = −.311), whereas vocabulary was not significantly correlated with breakdown fluency measures (r = −.229 for silent pause rate, r = −.227 for mean pause duration). Uchihara et al. (2019 2021) and Clenton et al. (2021) measured productive vocabulary knowledge via a word association task (Lex30; Meara and Fitzpatrick 2000) with Japanese adult learners of English as a foreign language. In line with previous studies using receptive vocabulary tests (e.g., de Jong and Mora 2019; Miralpeix and Muñoz 2018), productive vocabulary knowledge was significantly correlated with a range of fluency measures such as fluency judgement by trained raters (r = .342 in Uchihara and Saito 2019) and objective measures such as number of silent pauses (r = −.39 in Clenton et al. 2021) and articulation rate (r = .48 in Uchihara et al. 2021).

Studies have also investigated whether learners with larger vocabulary are able to produce more accurate and sophisticated language in terms of grammatical and lexical usage (Koizumi and In'nami 2013; Miralpeix and Muñoz 2018; Noreillie et al. 2020; Uchihara and Clenton 2020). Miralpeix and Muñoz (2018) measured receptive vocabulary knowledge of 42 Catalan/Spanish learners of English through a yes/no vocabulary recognition test. L2 speech elicited through a semi-guided interview was then rated by the researchers based on lexico-grammatical criteria. A significant and medium correlation between receptive vocabulary and lexico-grammatical features of spoken production (r = .413). Uchihara and Clenton (2020) examined the link between receptive vocabulary knowledge (via a checklist task) and lexical spoken proficiency using both human rating (lexical richness) and corpus-based measures (word frequency). Despite the significant correlation found for lexical rating (r = .552), learners' receptive knowledge did not significantly correlate with corpus-based frequency indices (r = .173 and .274), leading the authors to conclude that a larger vocabulary knowledge does not necessarily guarantee production of more sophisticated L2 words in spontaneous oral narrative. Noreillie et al. (2020) approached this issue more systematically with Flemish low-intermediate learners of French. The study elicited L2 speech through two dialogic speaking tasks and measured both receptive and productive vocabulary knowledge through multiple-choice and gap-filling tasks. Their findings partially support Uchihara and Clenton (2020) with a significant and smaller correlation between lexical rating and receptive vocabulary knowledge (r = .32). However, no significant link was found for productive vocabulary knowledge (r = .27). The study also indicated the importance of task effect as the significant correlation between vocabulary and lexical rating was not consistent across the two speaking tasks. A few other studies have reported mixed findings regarding the role of vocabulary knowledge in pronunciation accuracy. Using a yes/no vocabulary recognition test, Miralpeix and Muñoz (2018) found a significant and medium correlation

with pronunciation rating score (r = .311). In contrast, Uchihara and Saito (2019) found a negligible magnitude of association between productive vocabulary knowledge and accentedness rating (r = .034), and a slighter larger but statistically non-significant correlation for comprehensibility rating (r = .271). Mariano and Santiago (2020) found a medium correlation between vocabulary knowledge (DIALANG test score) and human ratings of foreign accentedness (r = .35), yet no significant correlations were observed for acoustic measures of pronunciation accuracy (p > .05).

Modality of vocabulary knowledge and speaking proficiency

Although a growing number of studies have documented the important relationship between vocabulary knowledge and L2 speaking proficiency (oral fluency, in particular), mixed findings regarding the strength of the relationship (r = .034 to .688) appear to leave us with some degree of uncertainty regarding how reliably and accurately vocabulary measures can predict learners' oral proficiency. Among the many factors contributing to the emerging inconsistency of results (e.g., different speaking tasks, participants' L2 proficiency, human rating vs. objective measures, sample size, measurement error), vocabulary test format can also be considered a major factor. The field of L2 vocabulary-speaking research has indeed progressed since Hilton's (2008) study by adopting different test formats including a wide array of receptive and productive vocabulary tasks. However, it should be noted that all the aforementioned studies (presented in Table 1) have measured vocabulary knowledge in written form. This obvious lack of attention to modality (written vs. spoken) is rather surprising given our earlier point that that incongruence in test format can lead to anomalous and inaccurate results. The degree to which L2 speaking research lags behind on this modality issue within the L2 vocabulary-proficiency research field is evident from the upsurge in the number of L2 listening studies calling for methodological improvement on the choice of test modality (McLean et al. 2015) and encouraging researchers to test vocabulary knowledge in spoken form (Hui and Godfroid 2020; Matthews and Cheng 2015; Vafaee and Suzuki 2020; Wallace 2020; Zhang and Zhang 2020). In particular, the lack of attention to test modality in L2 speaking research might account for inconsistent findings for the relationship between vocabulary and pronunciation (Mariano and Santiago 2020; Miralpeix and Muñoz 2018; Uchihara and Saito 2019). It is reasonable to assume that measurement of the spoken form of L2 words would better reflect how accurately learners can pronounce words and sentences. To the best of our knowledge, only two studies (Alhazmi and Milton 2016; Milton et al. 2010), have measured receptive vocabulary knowledge in both written and spoken modes using a yes/no vocabulary test format (X_Lex and Aural Lex). In these studies, spoken vocabulary recognition was more strongly associated with IELTS speaking scores (*rho* = .58 and .71) compared to written vocabulary recognition (*rho* = .42 and .35). Although revealing, such findings are limited to small sample sizes (N = 27 and 30) and receptive vocabulary measures (yes/no checklist tasks). Hence the need for more studies in this area with larger sample sizes and a greater variety of validated vocabulary test formats. The current study sets out to address these gaps by measuring productive vocabulary knowledge in written and spoken formats in an attempt to

determine the relationship between written and spoken vocabulary knowledge and by exploring the extent to which vocabulary is related to four aspects of L2 speaking proficiency (fluency, lexis, pronunciation, grammar). Our two research questions were as follows:

(1) To what extent is spoken and written productive vocabulary knowledge associated?
(2) To what extent does spoken productive vocabulary knowledge, in comparison to written productive vocabulary knowledge, predict different aspects of L2 speaking proficiency (fluency, lexis, pronunciation, grammar)?

Method

Participants

Forty-six undergraduate and postgraduate university L2 students (28 females, M_{age} = 29, *range* = 18 to 51) from 15 different countries participated in this study. Participants spoke different L1s including Japanese (*n* = 14), Chinese (*n* = 10), Kazakh (*n* = 5), Arabic (*n* = 4), Turkish (*n* = 3), Thai (*n* = 2), and other languages (*n* = 1 for Urdu, Malay, Portuguese, Farsi, Italian, Greek, Russian, and Spanish). Most of the participants were postgraduate students (except 2 undergraduates) at a UK university. In terms of L2 proficiency, participants were considered advanced L2 learners of English on the basis that they had lived in the UK for at least 7 months before the time of testing and every participant had achieved an equivalent score of at least IELTS 6.5 for admission to the university (the cohort had taken a variety of English language proficiency tests for admission, including: TOEFL, IELTS, as well as an in-house university entrance examination).

Productive vocabulary tests

This study adopted a word association elicitation task (Lex30; Meara and Fitzpatrick 2000) for assessing L2 productive vocabulary in written and spoken forms. Bearing in mind the limited repertoire of spoken productive vocabulary tests, we believe that using the spoken version of Lex30 is appropriate because it is the most extensively validated measure of productive vocabulary knowledge since the test was developed by Meara and Fitzpatrick (e.g., Clenton 2010; Fitzpatrick 2007; Fitzpatrick and Clenton 2010 2017; Fitzpatrick and Meara 2004; Walters 2012). Particularly relevant is part of the large-scale validation data in Fitzpatrick and Clenton (2010), which suggested that the spoken Lex30 has a great potential as a measure of spoken productive knowledge. The initial report showed that although there was no significant difference between the scores on the written and spoken versions of the test (t = 0.751, p = 0.457), a small correlation was observed (r = .391, p < .01), suggesting construct differences between written and spoken productive knowledge.

W_Lex30

Lex30 is a frequency-based vocabulary test developed based on the assumption that demonstrating knowledge of lower frequency words reflects a larger vocabulary size (Beglar 2010; Laufer and Nation,

1995, 1999; Webb et al. 2017). In order to distinguish the standard (written) version of the Lex30 from the spoken version of the test, we hereafter refer it to as W_Lex30 ('W' indicating written format). Assessment procedures adopted in this study followed previous Lex30 studies (see Meara and Fitzpatrick 2000, pp. 22-23; Fitzpatrick and Clenton 2010, p. 539). The Lex30 task presents L2 learners with 30 carefully selected cues (see Meara and Fitzpatrick 2000: 22, for their selection criteria), to which participants are asked to write up to four related words per each cue within 15 minutes (i.e., 30 seconds for each cue)—e.g., *attack* > *game*, *offense*, *defense*, *war* (see Appendix 1 for a list of the 30 cues). When participants did not produce four words within 30 seconds, they moved on to the next cue. Each set of responses amounts to a theoretical maximum of 120 words (up to 4 responses multiplied by the 30 cues). Prior to frequency profile analysis, each set of elicited responses was processed manually—i.e., misspellings were corrected, responses were lemmatized, and repetitions of the same responses were removed. Following Fitzpatrick and Meara (2004), processed responses were profiled according to lexical frequency information provided by JACET 8000 (JACET 2003).[1] Individual raw scores were then calculated per participant. All responses qualified as infrequent words (outside the first 1,000 frequency words) were awarded a point (proper nouns, numbers, and structure words were not included in scoring).

S_Lex30

For the spoken version of Lex30 we employed S_Lex30 ('S' indicating spoken format), originally developed by Fitzpatrick and Clenton (2010: 546-547). In the S_Lex30 test, all scoring procedures were the same as W_Lex30 with two exceptions. First, in order to avoid practice effect, we used a set of 30 cue words from Fitzpatrick and Clenton (2010), different from cue words used for W_Lex30 (see Appendix 2 for a list of the 30 cues). The cue words were selected according to the same criteria as Meara and Fitzpatrick (2000), with a different word list (JACET 8000 list; JACET 2003). Based on part of their validation data, Fitzpatrick and Clenton (2010: 542) concluded that the original list (used for W_Lex30) and the new list (used for S_Lex30) 'meet at least threshold criteria for equivalence' for research purposes (Hatch and Lazaraton, 1991).[2] Second, in the S_Lex30 task, participants were asked to read a cue word on a card and then verbalize up to four responses to each cue within 30 seconds. Once participants had produced four words for a cue or failed to do so within 30 seconds, the researcher showed the next card. As the test was administered individually, upon the completion of the S_Lex30 task, the researcher asked participants to clarify any incomprehensible responses (e.g., words that were pronounced too fast or not loud enough, or due to strong foreign accent). When participants could account for a word in some way (by repeating the word or spelling it out, indicating the meaning), it was included in the data set; when they could not, it was discarded. For this reason, the spoken vocabulary scores did not penalize non-nativelike pronunciation. Instead, we gave credit for partial word knowledge of spoken forms in order to make the scoring procedure comparable to that for W_Lex30 (i.e., minor misspellings were corrected; Meara and Fitzpatrick 2000). Participants' responses were audio recorded and later transcribed, and processed and scored in the same way as

for W_Lex30 data.

L2 speaking measures

We used a story-narrative task to elicit spontaneous L2 speech samples. The rationale behind our choice of task relates to storytelling and narrative accounting for a large proportion of conversation in daily life (Willis and Willis 2007) and being extensively used for research purposes (Skehan 2009). IELTS speaking band descriptors (accessible at https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en) were adopted to rate the speech samples (Isaacs et al. 2015) because they are widely recognized and popularly used for university admissions criteria and other academic purposes. The story-narrative task is based on a sequenced set of four picture prompts. The speech samples were then rated according to IELTS speaking descriptors consisting of four linguistic components (fluency, vocabulary, grammar, and pronunciation). In this study, speaking proficiency was operationalized based on human rating rather than objective measures (i.e., quantification of given linguistic features). The rationale behind the choice of rating measures echoes the language assessment literature arguing that the ultimate goal of teaching and testing L2 oral proficiency is how listeners perceive L2 speakers (Isaacs and Thomson 2013; Pallotti 2009). Extremely high scores derived from the quantification of linguistic features does not necessarily mean that the resultant performance is perceived as communicatively functional by listeners (for cases of lexical sophistication and fluency, see Saito et al. 2016 and Munro and Derwing 2001, respectively).[3]

Speech materials

The participants were asked to describe a four-strip cartoon (STEP 2015) immediately after studying the pictures to familiarize themselves with the story for approximately one minute. The cartoon prompt consisted of: (a) a couple finding two construction workers placing a signpost notifying of a new shopping mall to be constructed, the couple consider that they have to do something; (b) several days later the couple start a campaign and ask local people to sign a petition against the upcoming construction; (c) three months later the construction plan is cancelled; and (d) a month later the couple look upset to find a local newspaper headline saying unemployment rates are on the rise (see Appendix 3 for the picture prompt). All elicited speech samples were recorded individually in a quiet university office, and recorded digitally. The length of the speech samples varied (1–4 minutes) with no explicit time limit set in advance.

Speech ratings

Three L1 (English) speaking raters were recruited (2 females, 1 male) at a UK university. All of the raters were part of a university assessment team for an English language proficiency test, and had ample professional experiences of rating a variety of L2 learners' speech for high-stakes purposes (e.g., admission to undergraduate programmes). None of the raters reported any hearing difficulties. To rate each participant's spoken linguistic features, a public version of the IELTS speaking band

descriptors was used. IELTS speaking descriptors contain 4 linguistic components: (a) Fluency and coherence, (b) Lexical resources, (c) Grammatical range and accuracy, and (d) Pronunciation. To rate all speech samples, the raters were first given a few minutes to study IELTS speaking descriptors and then rated a training set of four speech samples to familiarize themselves with what was required. Subsequently, the raters listened to speech samples from each participant in a random order. While listening to one sample at a time from the beginning to the end, each of the raters assigned four separate ratings (from 1 to 9) based on the four IELTS speaking descriptors. An inter-rater consistency for all aspects—fluency (α = .84), vocabulary (α = .77), grammar (α = .83), and pronunciation (α = .76)—were considered acceptable for research purposes exceeding a minimum benchmark value (α = .70-.80; Larson-Hall 2010).

Procedure

We conducted testing sessions individually, in one sitting per participant with one researcher (the first author). The participants were first required to complete consent forms and a language background questionnaire (regarding age, gender, L1, length of residence in the UK, level of education). Participants then completed the written form of Lex30 (W_Lex30), the spoken form of Lex30 (S_Lex30), and the speaking task (the cartoon-strip description). Prior to taking each of the Lex30 tests, the participants completed a practice task with a training set of example cue words. They were instructed to write or verbalize up to four words that came to mind for each cue without any other instruction or feedback provided. For the speaking task, they were instructed to describe a four-strip cartoon immediately after studying the pictures to familiarize themselves with the story for one minute. No time limitation was set in advance so that speakers did not feel anxious, and they were encouraged to speak as much as needed to describe the given pictures (Saito et al. 2016). Participants knew that their speech performance would be assessed, but were not informed of how it would be assessed. Participants were allowed to look at the cartoon while describing it and given as much time as they wanted for the task. Subsequent to administration of the three tests to all participants, the speech samples were evaluated by three raters, based on IELTS speaking descriptors. Through an individual appointment made by the researcher with each of the three raters, they were asked to rate speech samples in the researcher's presence. To avoid fatigue effects, raters evaluated 46 samples in two or three sessions. Each rater took approximately three hours for each rating.

Data analysis

Prior to conducting statistical analyses to answer the two research questions, preliminary assumptions for parametric tests were checked. The Shapiro-Wilk tests of normality indicated that all data except two speech ratings (p = .049 for vocabulary, p = .017 for grammar) were normally distributed. The normal distribution of vocabulary and grammar ratings was further confirmed by an examination of the histograms and skewness statistics (absolute values of skewness statistics for the two scores were less than 1.0, Larson-Hall 2010). Descriptive statistics for the scores of vocabulary

tests and speech ratings are presented in Tables 2 and 3, respectively. Because IELTS band descriptors are not intended to be used to assess speech performance elicited via a picture narrative task, an exploration of the rating data was needed to verify whether the rating scales worked properly in our data set. We therefore conducted a many-facet Rasch analysis and confirmed that our raters had appropriately assessed L2 speech samples using IELTS descriptors (see Supplementary Material for the detail of the results).

Table 2. *W_Lex30 and S_Lex30 task scores* (*N* = 46)

	M	*SD*	*Min*	*Max*	[95% confidence interval]
W_Lex30	43.5	9.7	28	70	[40.64, 46.41]
S_Lex30	42.7	10.7	21	62	[39.53, 45.86]

Note. Maximum scores for W_Lex30 and S_Lex30 were 120.

Table 3. *Means, standard deviations, range, and reliability of speaking ratings* (*N* = 46)

	M	*SD*	Range	Reliability
Fluency	5.8	0.8	4.0-7.7	0.84
Vocabulary	5.9	0.8	4.3-7.3	0.77
Grammar	5.8	0.8	4.7-8.0	0.83
Pronunciation	6.0	0.7	4.7-7.7	0.76

To answer the first research question regarding the written and spoken vocabulary knowledge relationship, Pearson correlation analysis and a paired-samples *t*-test were conducted between W_Lex30 and S_Lex30 scores. To answer the second research question regarding the relationship between productive vocabulary knowledge and L2 speaking proficiency, a series of Pearson correlation analysis were conducted between two vocabulary measures (W_Lex30 and S_Lex30) and four speech ratings (fluency, vocabulary, grammar, pronunciation). In order to examine the unique contribution of spoken vocabulary knowledge while the effect of written vocabulary knowledge was controlled for, standard multiple regression analyses were conducted with S_Lex30 as the predictor, W_Lex30 as the covariate, and each of the four speech ratings as the outcome variable separately, resulting in four regression models. Statistical assumptions—linearity, collinearity, and homogeneity of variances—were checked before analyses were conducted. The size of correlation coefficient was interpretated according to the field-specific benchmarks of the effect size (small = .25, medium = .40, large = .60) provided by Plonsky and Oswald (2014).

Results

To what extent are spoken and written vocabulary knowledge associated?

The result of a paired-samples *t*-test showed no significant difference between W_Lex30 and S_Lex30 scores, $t = 0.66$, 95% CI [−1.70, 3.35], $p = .514$, $d = 0.10$, indicating that learners' productive vocabulary knowledge was consistent between the two modes. The result of correlation analysis showed a

significant correlation (r = .655, p < .001) between W_Lex30 and S_Lex30 scores, indicating that learners' written and spoken productive vocabulary was related. However, visual examination of the relationship presented in Figure 1 showed that some learners demonstrated a large gap between the two types of knowledge, indicating that the two versions of Lex30 scores are not perfectly matched.

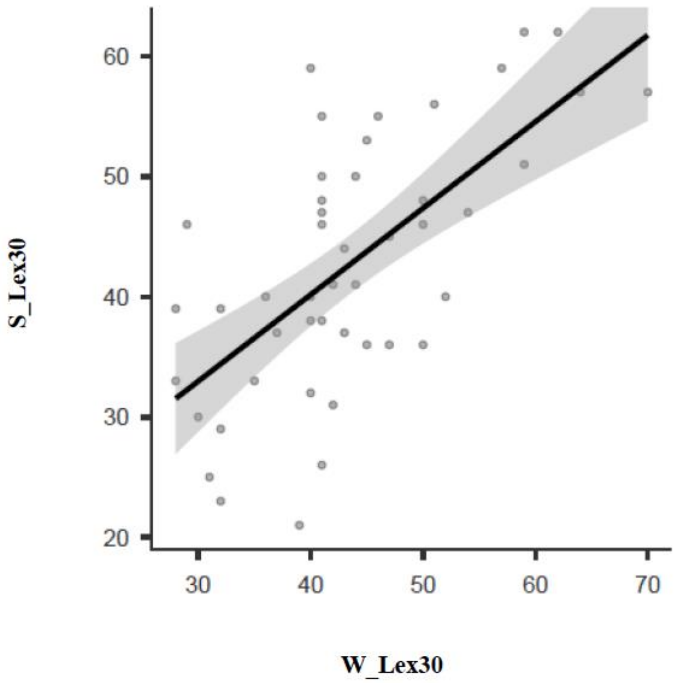

Figure 1. Scatterplot representing the relationship between W_Lex30 and S_Lex30 scores.

To what extent does spoken vocabulary knowledge predict L2 speaking proficiency?

The results of correlation analyses (Table 4) showed no significant correlations between the written productive vocabulary (W_Lex30) and L2 speech ratings. However, the spoken productive vocabulary (S_Lex30) significantly correlated with all L2 speech ratings except fluency. In order to further scrutinize the unique contribution of spoken vocabulary while the effect of written vocabulary was controlled, a series of multiple regression analysis was conducted for each of three speech ratings (vocabulary, grammar, pronunciation). The results showed that the relationship between the spoken vocabulary (S_Lex30) and pronunciation rating remained significant after the influence of the written vocabulary (W_Lex30) was accounted for (see Table 5). The correlation between S_Lex30 score and pronunciation rating remained a medium effect (partial correlation = .431) after the influence of W_Lex30 was statistically controlled for.

Table 4. *Correlations between W_Lex30 and S_Lex30 scores and L2 speech ratings*

	W_Lex30		S_Lex30	
	r	*p*	*r*	*p*
Fluency	.173	.251	.255	.087
Vocabulary	.201	.181	.314*	.034
Grammar	.264	.076	.327*	.027
Pronunciation	.273	.066	.492**	< .001

Note. * indicates $p < .05$ and ** indicates $p < .01$.

Table 5. *Regression analysis with W_Lex30 and S_Lex30 scores as the predictors and vocabulary, grammar, and pronunciation ratings as the outcome variables*

		95% Confidence Interval			
	β	Lower	Upper	*t*	*p*
Vocabulary					
S_Lex30	0.32	−0.07	0.71	1.67	.103
W_Lex30	−0.01	−0.40	0.39	−0.04	.966
Grammar					
S_Lex30	0.27	−0.12	0.65	1.41	.165
W_Lex30	0.09	−0.30	0.47	0.46	.645
Pronunciation					
S_Lex30	0.55	0.20	0.90	3.14	.003
W_Lex30	−0.09	−0.44	0.27	−0.49	.625

Discussion

The modality effect in testing vocabulary knowledge has been overlooked in the past few decades despite the increase in the number of studies exploring the role of vocabulary in L2 speaking. The current study can therefore be viewed as a response to this lack of attention by measuring the spoken form of productive vocabulary in order to provide further insight into the relationship between vocabulary and speaking proficiency. The general picture emerging from this study is in line with previous studies suggesting that (a) spoken and written vocabulary knowledge are related but independent constructs and (b) spoken vocabulary knowledge is closely associated with L2 speaking proficiency. More detailed discussion will follow in response to each of the two research questions.

In answer to the first research question, the scores of productive vocabulary knowledge measured via W_Lex30 and S_Lex30 were significantly correlated ($r = .655$) to a similar degree compared with the correlation ($r = .688$) reported for receptive vocabulary measures (Milton and Hopkins 2006). The large effect observed in this study compared to the correlation ($r = .391$) in Fitzpatrick and Clenton (2010) might be attributed to the different test-taking environments in the latter study. In Fitzpatrick and Clenton's study W_Lex30 and S_Lex30 tasks were administered in the classroom and laboratory settings respectively, whereas in the current study learners took the two tests in the same environment (for a further discussion of this issue, see Fitzpatrick and Clenton 2010, pp. 546–547). Furthermore, the result of an independent samples *t*-test showed no significant

difference between W_Lex30 and S_Lex30 scores at the group level. Yet, a further examination of the data revealed a large gap between the spoken and written vocabulary at the individual level. The lack of difference at the group level contrasts with the findings of earlier studies focusing on receptive knowledge, which suggested that spoken vocabulary is smaller than the written vocabulary (Milton and Hopkins 2006; Uchihara and Harada 2018). The inconsistent findings between receptive and productive vocabulary might be because the rate of development at the productive level might be more susceptible to individual differences. One possible factor might also be a learner's L1. Milton and Hopkins (2006) found that Arabic learners of English, who are more likely to experience difficulties with the orthography of English, recognized a smaller number of words in written mode than in spoken mode. Although our data of mixed L1 backgrounds did not allow further exploration of the L1 influence, three Arabic L1 learners did not appear to align with such an expected pattern (W_Lex30 > S_Lex30: 46 vs. 55, 41 vs. 47, 41 vs. 26). The tentative account of attributing the inconsistent findings to individual differences is also supported by L2 speech literature (Sakai and Moorman 2018). Speech perception training involving repeated exposure to target L2 sounds has been found to result in relatively consistent improvement on the ability to recognize L2 phonological forms. However, the degree of improvement on recognition accuracy may not necessarily lead to the improvement on the ability to produce L2 sounds accurately (Nagle 2017). Achieving advanced production accuracy requires strong motivation (Moyer 2014) and special language learning abilities such as phonemic coding (Granena and Long 2013).

In answer to the second research question, the results showed that the spoken vocabulary knowledge (measured via S_Lex30) was significantly correlated with three speech ratings (vocabulary, grammar, pronunciation), whereas no significant correlations were found between the written vocabulary knowledge (measured via W_Lex30) and all speech ratings. The follow-up regression analysis showed that the relatively strong link between spoken vocabulary knowledge and the pronunciation rating persisted after the effect of written vocabulary knowledge was statistically controlled for. These findings suggest the important role of modality (spoken vs. written) such that knowledge of the spoken form of words is more closely associated with L2 speaking proficiency. The effect of modality becomes more salient particularly when we looked into the relationship for pronunciation rating. Notably, the effect size for pronunciation observed in this study (r = .491) appears to be larger compared to the findings of earlier studies comparing pronunciation and written vocabulary knowledge (r = .311 in Miralpeix and Muñoz 2018; r = .35 in Mariano and Santiago 2020; r = .271 and .034 in Uchihara and Saito 2019). The relatively stronger correlation found in this study indicates that the strength of the vocabulary-speaking link depends on the mode in which vocabulary knowledge is tested as well as the aspect of speaking proficiency measured. Regarding the vocabulary and grammar ratings (r = .314 and .327), testing the spoken forms instead of written forms of words did not seem to enhance the strength of the vocabulary-speaking link compared to the findings of earlier studies using written vocabulary tests (r = .413 in Miralpeix and Muñoz 2018; r = .32 in Noreillie et al. 2020; r = .552 in Uchihara and Clenton 2020). It is possible that the mode difference, spoken or

written, may not have a significant impact on the predictive role of vocabulary for some aspects of L2 speech.

Counter to our expectation, neither W_Lex30 nor S_Lex30 scores significantly correlated with the fluency rating. The results were rather surprising given that the central role of vocabulary in oral fluency has been empirically documented in earlier studies (e.g., Clenton et al. 2021; de Jong and Mora 2019; Hilton 2008; Koizumi and In'nami 2013; Uchihara and Saito 2019). One possible reason might be due to differences in participants' L2 proficiency. The current study investigated advanced L2 learners, most of whom were postgraduate students and had lived in the English-speaking country for at least 7 months, whereas most of the studies reviewed above investigated learners in foreign language settings with an L2 proficiency ranging from beginner to intermediate. Another potential factor might relate to our choice of the rubric for fluency rating. In this study, rater judgements were not derived from purely fluency-focused rubric (i.e., fluency *and* coherence), which might have reflected learner discourse competence—the ability to comply with the expected text structure—as well as spoken fluency (Iwashita and Vasquez 2015). Given the possibility that different descriptions of fluency direct rater's attention to different linguistic features (Suzuki et al. 2021), this unique characteristic of the rated scores, unlike fluency measures adopted in other studies (e.g., optimal speech rate; Uchihara and Saito 2019), might have confounded the relationship between productive vocabulary knowledge and our fluency rating.

Conclusion

The findings of this study have revealed the related but partially independent constructs of written and spoken vocabulary knowledge at the productive level, and the important role of vocabulary knowledge in L2 speaking proficiency. The current study contributes to the ongoing debate about the central role of modality in testing vocabulary knowledge by demonstrating the major impact of modality difference on how predictive learners' vocabulary knowledge is of different aspects of L2 speech. If the purpose of testing vocabulary is to obtain an indication of L2 speaking proficiency, researchers and teachers should administer the test in spoken format. In particular, the use of written vocabulary measures is likely to underestimate L2 pronunciation proficiency. Since the ability to pronounce L2 forms accurately serves as a reliable predictor of general speaking proficiency (Iwashita et al. 2008; Saito et al. 2016), test users measuring the written form of L2 words might lose the key linguistic information necessary for gaining the general picture of learners' L2 speaking proficiency. The current study also suggests that the spoken version of the Lex30 task can serve as a practical tool for language teachers to quickly gauge vocabulary knowledge pertaining to L2 speaking proficiency. The Lex30 spoken task can be particularly useful when time and resources are limited for administering a full-scale speaking assessment. Although we need to use the test carefully in relation to the decision we try to make, the findings of the current study at least suggest that this test format could be a useful addition to the existing battery of vocabulary tests.

The current study has several limitations that further research can expand on in order to

explore the relationship between vocabulary knowledge and L2 speaking proficiency. First, to avoid test effects, the current study used two different sets of 30 cues to elicit L2 word responses in written and spoken form. Although the minimal threshold equivalence of the two sets of cue words was established in a previous validation study (Fitzpatrick and Clenton 2010), the results of the current study should be interpreted cautiously with this limitation in mind. One way to overcome this challenge is to use the same cue words and alternate administration of the two versions of the test to mitigate the effect of the test order. Ideally, some distractor tasks should intervene between the two tests. Another is to leave an interval between the two test-taking times in order to reduce the effect of taking the first test on the subsequent test-taking behavior (Fitzpatrick and Clenton 2010). Second, the current study adopted a global approach to scoring the spoken form of responses elicited through S_Lex30 task. More detailed analysis for scoring of the elicited responses is possible. Although this is beyond the scope of the current study, the spoken forms of words can be evaluated acoustically in terms of segmental and suprasegmental accuracy (Saito and Plonsky 2019). Finally, speaking performance was elicited using a single task. The findings of the relationship between vocabulary knowledge and oral proficiency based on a cartoon narrative task may not be generalizable to other contexts where different speaking tasks are used to elicit L2 speech. One direction for future research is to investigate the extent to which the vocabulary-and-speaking link is consistent across different speaking tasks with varying degrees of task complexity, topic familiarity, and planning time.

Notes

1. Although alternative updated word lists could be selected, this study used JACET 8000 (2003) given the accumulated evidence suggesting that this word list works for measuring productive vocabulary knowledge elicited through the Lex30 task (Fitzpatrick and Clenton 2010; Fitzpatrick and Meara 2004) and for the purpose of exploring the vocabulary-speaking relationship (Uchihara and Saito 2019). Future research might consider other word lists such as Nation's (2012) BNC/COCA word-family lists.

2. The current study used different cue items from different wordlists: items from Nation (1984) for W_Lex30 and items from JACET (2003) for S_Lex30. Fitzpatrick and Clenton (2010, p. 542) reported (a) a significant correlation (r = .692, p < .001) between the two parallel forms of the test, (b) no significant difference for the mean scores (t = 0.81, p = .425), and (c) no significant difference for the variance statistics (p = .277). Despite the relatively smaller correlation for the purpose of establishing ideal test equivalence, the significant correlation together with the results of the mean and variance statistics suggested that 'the Lex30 tests meet at least threshold criteria for equivalence'.

3. Ideally, both subjective and objective measures were used. However, given our wider focus in measuring speaking proficiency, it was not feasible to adopt objective measures to quantify all temporal, lexical, grammatical, and phonological features of L2 speech. Future research should narrow the scope of speaking proficiency and measure not only perceived proficiency using human rating but also linguistic features in greater detail such as lexical use (e.g., sophistication, diversity, density, and accuracy; Zaytseva et al. 2021),

phonological accuracy (e.g., acoustic measures of segmental and prosodic features; Saito and Plonsky 2019), and temporal properties (e.g., articulation rate, location of pauses; Suzuki et al. 2021).

References

Alhazmi, K., and J. Milton. (2016). Phonological vocabulary size, orthographic vocabulary size, and EFL reading ability among native Arabic speakers. *Journal of Applied Linguistics*, *30* 26–43.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing*, *18*(3), 191–208. DOI: 10.1016/j.jslw.2009.05.003

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing* 2*7*, 101–108. DOI: 10.1177/0265532209340194

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

Clenton, J. (2010). I*nvestigating the construct of productive vocabulary knowledge with Lex30. Unpublished doctoral dissertation*. University of Swansea, Swansea, UK.

Clenton, J., de Jong, N. H., Clingwall, D., and Fraser, S. (2021). Investigating the extent to which vocabulary knowledge and skills can predict aspects of fluency for a small group of pre-intermediate Japanese L1 users of English (L2). In J. Clenton and P. Booth, *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 126–145). Abingdon, UK: Routledge.

de Jong, N. H., and Mora, J. C. (2019). Does having good articulatory skills lead to more fluent speech in first and second languages? *Studies in Second Language Acquisition*, *41*(1) 227–239. DOI: 10.1017/S0272263117000389

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., and Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34. DOI: 10.1017/S0272263111000489

Fitzpatrick, T. (2007). Productive vocabulary tests and the search for concurrent validity. In H. Daller, J. Milton and J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 116-132). Cambridge University Press.

Fitzpatrick, T., and Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing* 2*7*(4), 537–554. DOI: 10.1177/0265532209354771

Fitzpatrick, T., and Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly*, *51*(4), 844–867. DOI: 10.1002/tesq.356

Fitzpatrick, T., and Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *Vigo international journal of applied linguistics*, *1*, 55–74.

Granena, G., and Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains: *Second Language Research* 2*9*(3), 311–343. DOI: 10.1177/0267658312461497

Harrington, M., and Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, *37*(4), 614–626. DOI: 10.1016/j.system.2009.09.006

Hatch, E., and Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Newbury House.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal*, *36*(2), 153–166. DOI: 10.1080/09571730802389983

Hui, B., and Godfroid, A. (2020). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, 1–27. DOI: 10.1017/S0142716420000193

Isaacs, T., and Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. DOI: 10.1080/15434303.2013.769545

Isaacs, T., Trofimovich, P., Yu, G., and Chereau, M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS Pronunciation scale. *IELTS Research Reports Online Series No.4*.

Iwashita, N., Brown, A., McNamara, T., and O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29(1) 24–49. DOI: 10.1093/applin/amm017

Iwashita, N., and Vasquez, C. (2015). An examination of discourse competence at different proficiency levels in IELTS Speaking Part 2. *IELTS Research Reports Online 5*, 1–44.

Japan Association of College English Teachers (JACET) Basic Word Revision Committee (Ed.). (2003). *JACET List of 8000 Basic Words*. Tokyo: Author.

Jelani, N. A. M., and Boers, F. (2018). Examining incidental vocabulary acquisition from captioned video: Does test modality matter? *ITL-International Journal of Applied Linguistics*, *169*(1), 169–190. DOI: 10.1075/itl.00011.jel

Jeon, E. H., and Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. DOI: 10.1111/lang.12034

Koizumi, R., and In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research, 4,* 900–913. DOI: 10.4304/jltr.4.5.900-913

Kormos, J. (2006). *Speech production and second language acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 454–476). Routledge. https://doi.org/10.4324/9780429291586-24

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Laufer, B., and Aviad–Levitzky, T. A. M. I. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, *101*(4), 729–741. DOI: 10.1111/modl.12431

Laufer, B., and Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. DOI: 10.1093/applin/16.3.307

Laufer, B., and Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. DOI: 10.1177/026553229901600103

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* MIT Press.

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

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

Mairano, P., and Santiago, F. (2020). What vocabulary size tells us about pronunciation skills: Issues in assessing L2 learners. *Journal of French Language Studies*, *30*(2), 141–160. DOI: 10.1017/S0959269520000010

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

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, and J. Williams (Eds.), *Performance and Competence in Second Language Acquisition* (pp. 35–53). Cambridge University Press.

Meara, P., and Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System* 2*8*, 19–30. DOI: 10.1016/S0346-251X(99)00058-5

Meara, P., and Miralpeix, I. (2016). *Tools for researching vocabulary*. Multilingual Matters.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Vedder, I. Bartning, and M. Martin (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232). Second Language Acquisition and Testing in Europe Monograph Series 1.

Milton, J., and Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, *63*(1), 127–147. DOI: 10.1353/cml.2006.0048

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

Miralpeix, I., and Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*(1), 1–24. DOI: 10.1515/iral-2017-0016

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, *35*(4), 418–440. DOI: 10.1093/applin/amu012

Munro, M. J., and Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition* 2*3*(4), 451–468. DOI: 10.1017/S0272263101004016

Nagle, C. L. (2017). Examining the temporal structure of the perception-production link in second language acquisition: A longitudinal study. *Language Learning*, *68*(1) 234–270. DOI: 10.1111/lang.12275

Nation, I. S. P. (1984). Vocabulary lists. Victoria University of Wellington, English Language Institute, Wellington, New Zealand.

Nation, I. S. P. (2012). *The BNC/COCA word family lists.* https://www.wgtn.ac.nz/lals/about/staff/paul-nation

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

Noreillie, A. S., Desmet, P., and Peters, E. (2020). Factors predicting low-intermediate French learners' vocabulary use in speaking tasks. *Canadian Modern Language Review*, *76*(3), 194–217. DOI: 10.3138/cmlr-2019-0018

Noreillie, A. S., Kestemont, B., Heylen, K., Desmet, P., and Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages: An approximate replication study of Stæhr (2009). *ITL-International Journal of Applied Linguistics*, *169*(1) 212–231. DOI: 10.1075/itl.00013.nor

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. DOI: 10.1093/applin/amp045

Plonsky, L., and Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. DOI: 10.1111/lang.12079

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*(2) 282–307. DOI: 10.3138/cmlr.56.2.282

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

Qian, D. D., and Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 66–80). Routledge.

Saito, K., and Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. DOI: 10.1111/lang.12345

Saito, K., Trofimovich, P., and Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*(2) 217–240. DOI: 10.1017/S0142716414000502

Sakai, M., and Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187–225. DOI: 10.1017/S0142716417000418

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. DOI: 10.1093/applin/amp047

Society for Testing English Proficiency (STEP). (2015) *Eiken jyunikkyu kako 6kai zen mondaisyu [A collection of past 6 administration of the exam for STEP test grade pre-1]*. Obunsha.

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

Suzuki, S., Kormos, J., and Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*. DOI: 10.1111/modl.12706

Uchihara, T. (2021). Vocabulary and speaking: Current research, tools, and practices. In J. Clenton and P. Booth, *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 121–125). Routledge.

Uchihara, T. (2022). Is it possible to measure word-level comprehensibility and accentedness as independent constructs of pronunciation knowledge? *Research Methods in Applied Linguistics*. Advance online publication. https://doi.org/10.1016/j.rmal.2022.100011

Uchihara, T., and Clenton, J. (2020). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research* 24(4), 540–556. DOI: 10.1177/1362168818799371

Uchihara, T., and 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. DOI: 10.1002/tesq.453

Uchihara, T., and Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, *47*(1), 64–75. DOI: 10.1080/09571736.2016.1191527

Uchihara, T., Saito., K., and Clenton, J. (2021). Reexamining the relationship between productive vocabulary knowledge and second language oral ability. In J. Clenton and P. Booth, *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 146–165). Routledge.

Vafaee, P., and 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. DOI: 10.1017/S0272263119000676

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

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*. DOI: 10.1111/lang.12424

Walters, J. (2012). Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly*, *9*(2), 172–185. DOI: 10.1080/15434303.2011.625579

Webb, S., Sasao, Y., and Oliver, B. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Willis, D., and Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.

Yanagisawa, A., and Webb, S. (2020). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371–386). Routledge.

Zareva, A., Schwanenflugel, P., and Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition* 27, 567–595. https://doi.org/10.1017/S02722631050254

Zaytseva, V., Miralpeix, I., and Pérez-Vidal, C. (2021). Because words matter: Investigating vocabulary development across contexts and modalities. *Language Teaching Research* 25(2), 162–184. https://doi.org/10.1177/1362168819852976

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

Appendix 1: Task format and cue words – Written Version of Lex30 (W_Lex30)

For each word, write up to four other words it makes you think of:

attack				
board				
close				
cloth				
dig				
dirty				
disease				
experience				
fruit				
furniture				
habit				
hold				
hope				
kick				
map				
obey				
pot				
potato				
real				
rest				
rice				
science				
seat				
spell				
substance				
stupid				
television				
tooth				
trade				
window				

Appendix 2: Cue Words – Spoken version of Lex30 (S_Lex30)

1. away
2. blow
3. brush
4. chance
5. common
6. dance
7. district
8. ever
9. famous
10. flag
11. get
12. head
13. insect
14. knee
15. list
16. mat
17. mountain
18. oil
19. pattern
20. policeman
21. public
22. religion
23. secret
24. shirt
25. sorry
26. smell
27. spirit
28. surprise
29. telephone
30. tool

Appendix 3: A speech material (i.e., a four-strip cartoon) from a published past exam questions of the STEP test for Grade pre-1 (STEP 2015, p. 120)