Bilingualism: Language and Cognition

cambridge.org/bil

Research Article

Cite this article: Lee, S.T., van Heuven, W.J.B., Price, J.M. and Leong, C.X.R. (2025). Assessing bilingual language proficiency with a yes/no vocabulary test: the role of form-meaning vocabulary knowledge. *Bilingualism: Language and Cognition*, 1–17 https://doi.org/10.1017/S1366728924001007

Received: 06 September 2023
Revised: 07 November 2024
Accepted: 09 November 2024

Keywords:
L1 proficiency; L2 proficiency; LexMAL; vocabulary knowledge; yes/no vocabulary test

Corresponding author:
Soon Tat Lee;
Email: soontat.lee@nottingham.edu.my

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

Assessing bilingual language proficiency with a yes/no vocabulary test: the role of form-meaning vocabulary knowledge

Soon Tat Lee[1], Walter J. B. van Heuven[2], Jessica M. Price[1] and Christine X. R. Leong[1]

[1]School of Psychology, University of Nottingham, Selangor, Malaysia and [2]School of Psychology, University of Nottingham, Nottingham, UK

Abstract

Validated yes/no vocabulary tests that measure bilinguals' language proficiency based on vocabulary knowledge have been widely used in psycholinguistic research. However, it is unclear what aspects of test takers' vocabulary knowledge are employed in these tests, which makes the interpretation of their scores problematic. The present study investigated the contribution of bilinguals' form-meaning knowledge to their item accuracy on a Malay yes/no vocabulary test. Word knowledge of Malay first- ($N = 80$) and second-language ($N = 80$) speakers were assessed using yes/no, meaning recognition, form recognition, meaning recall and form recall tests. The findings revealed that 59% of the variance in the yes/no vocabulary test score was explained by the accuracy of the meaning recognition, form recognition and meaning recall tests. Importantly, the item analysis indicated that yes/no vocabulary tests assess primarily knowledge of form recognition, supporting its use as a lexical proficiency measure to estimate bilinguals' receptive language proficiency.

Highlights

The core findings of the paper revealed that:

- Yes/no vocabulary tests measure form-meaning knowledge.
- Yes/no vocabulary tests are likely to require form recognition knowledge.
- Yes/no vocabulary tests are less likely to require meaning/form recall knowledge.

1. Introduction

Language proficiency plays an important role in our understanding of bilingual language processing (e.g., Fromont et al., 2020; Sarrett et al., 2022; Singh et al., 2022; Tosun & Filipović, 2022). Despite often being construed as a moderating variable in bilingual research, there is great variability in how language proficiency is operationalized and measured (Park et al., 2022; Puig-Mayenco et al., 2023; Surrain & Luk, 2019; Treffers-Daller, 2019; Tremblay, 2011). For instance, language proficiency measured by objective measures can be operationalized based on different language constructs, such as vocabulary knowledge or morphosyntactic knowledge (Treffers-Daller, 2019). In these measures, test takers' language proficiency is often expressed as a score on a scale (e.g., in percentage) and is interpreted based on the construct that the test purports to measure (Hulstijn, 2012). For example, a test taker who receives a higher score in a vocabulary-based language proficiency test is assumed to have a higher level of language proficiency than a test taker who receives a lower score. Interpreting language proficiency measures depends on how language proficiency is conceptualized in the tests, including the purpose of the test, target learners, context of testing and the aspects and levels of language constructs in consideration (Schmitt et al., 2019). These specifications regarding a test should be established before and during test development and validation, and researchers should select the language proficiency measure that matches the experimental context so that test scores meaningfully inform the language ability that the experiment aims to investigate. As a result, accurately conceptualized tests could improve the conclusions made about the relationship between language ability and language processing (Mainz et al., 2017).

Among all types of language tests, vocabulary tests have been widely used as an objective estimate of language proficiency in research, because vocabulary knowledge is one of the fundamental constructs that underlie language proficiency (Brysbaert et al., 2017; Nation & Beglar, 2007; Qian & Lin, 2020; Schmitt et al., 2015). Moreover, the ability to recognize word forms and access their meanings in the mental lexicon is crucial for reading comprehension (Harrington, 2018). As the initial stage of reading, word recognition (i.e., access to form

CAMBRIDGE
UNIVERSITY PRESS

and meaning knowledge) is a strong predictor of L1 (e.g., Holmes, 2009) and L2 reading comprehension (e.g., Jeon & Yamashita, 2014). Therefore, in the field of psycholinguistics, yes/no vocabulary tests such as the Lexical Test for Advanced Learners of English (LexTALE, Lemhöfer & Broersma, 2012) and its extensions in French (LexTALE-FR: Brysbaert, 2013), Spanish (Lextale-Esp: Izura et al., 2014), Chinese (LEXTALE_CH: Chan & Chang, 2018; LexCHI: Wen et al., 2023), Italian (LexITA: Amenta et al., 2021), Portuguese (LextPT: Zhou & Li, 2022), Finnish (Lexize: Salmela et al., 2021) and Malay (the Lexical Test for Malay Speakers, LexMAL, Lee et al., 2023) have been used to estimate bilinguals' language proficiency. These tests are freely available and are time-efficient, allowing a relatively large number of words to be tested in 5 min. Positive correlations were found between yes/no vocabulary test scores and other language proficiency measures such as Quick Placement Test (Lemhöfer & Broersma, 2012; Masrai, 2022) and translation tasks (Lee et al., 2023; Lemhöfer & Broersma, 2012; Wen et al., 2023), demonstrating the validity of the tests as measures of language proficiency. Furthermore, yes/no vocabulary test scores have been shown to predict language performance in other language tasks such as lexical decision and visual word recognition (e.g., Diependaele et al., 2013; Lemhöfer & Broersma, 2012; Wen & van Heuven, 2017). In addition, these tests can be used to discriminate L1 and L2 speakers by grouping test takers into higher and lower proficiency groups based on their scores (e.g., Brysbaert, 2013; Izura et al., 2014; Lee et al., 2023; Wen et al., 2023). This is useful for research studying language proficiency effects (e.g., comparing performance of L1 and L2 speakers) or language processing across speaker groups of the same language.

Although the validity of yes/no vocabulary tests has been consistently demonstrated in past studies (e.g., Lee et al., 2023; Lemhöfer & Broersma, 2012; Masrai, 2022; Wen et al., 2023; Zhang et al., 2020), it is unclear precisely which aspects of test takers' vocabulary knowledge are assessed in these tests, making meaningful score interpretation problematic. Furthermore, different test instructions were used by different yes/no vocabulary tests, which makes test score interpretation even more difficult. For instance, some yes/no vocabulary tests instruct test takers to indicate "yes" when they "know" the meaning of the target words (e.g., V_YesNo: Meara & Miralpeix, 2016), whereas other yes/no vocabulary tests employ an unspeeded lexical decision format, in which test takers are required to decide whether the letter strings presented are real words (e.g., LexTALE: Lemhöfer & Broersma, 2012). The commonality these tests share is the lack of direct demonstration of word knowledge during performance. Therefore, a "yes" response in the tests may reflect word knowledge that ranges from being able to recognize the meaning and/or word form to being able to produce it. As knowing a word involves knowledge of different word aspects that can be known to different levels of strength (Nation, 2020; Qian & Lin, 2020), it is unclear to what extent participants could recognize and produce the word forms or meanings when they correctly indicate a "yes" response in the yes/no vocabulary test. To this end, the present study aimed to investigate the role of vocabulary knowledge in bilinguals' performance in a yes/no vocabulary test.

Table 1. Nation's (2013) framework of the components involved in knowing a word

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

Note. R = receptive knowledge; P = productive knowledge. Adapted from Nation (2013).

Vocabulary knowledge is a multifaceted unidimensional construct that contains several interrelated but distinct aspects of word knowledge (González-Fernández, 2022; González-Fernández & Schmitt, 2020; Schmitt, 2010). According to Nation (2013, 2020, 2022), mastery of nine aspects of word knowledge is required to achieve lexical proficiency and each can be divided into receptive and productive knowledge (see Table 1). The receptive/productive conceptualization entails how various word knowledge aspects are used for communicative purposes in real life. Receptive knowledge refers to the skills needed to recognize and understand a lexical item well enough to extract communicative meaning from speech or writing, whereas productive knowledge involves the skills of recalling and producing a lexical item to encode communicative content in speech or writing (González-Fernández & Schmitt, 2020; Nation, 2020; Schmitt, 2010). The different aspects of word knowledge (receptive and productive) have different difficulty levels and can be mastered to various degrees at different stages of word acquisition (González-Fernández & Schmitt, 2020; Nation, 2020). For instance, the knowledge of form-meaning connections (e.g., recognizing "table" as a word form for the furniture with a flat top and one or more legs) is one of the fundamental aspects in initial vocabulary learning, whereas other aspects of word knowledge (e.g., a constraint on the use of word forms) slowly build up as proficiency develops. Therefore, examining the interrelations between these word knowledge aspects may help to understand their unique contribution to overall lexical proficiency.

It is, however, difficult to truly measure distinct word knowledge aspects in isolation based on the skill-based receptive/productive definitions (Schmitt, 2010). Alternatively, prior studies tapped into receptive/productive word knowledge by adopting standard recognition and recall test formats. Recognition and recall of word knowledge are commonly assessed in vocabulary tests to gain insights into the strength of receptive and productive vocabulary knowledge (e.g., González-Fernández & Schmitt, 2020; Laufer &

Goldstein, 2004). A word recognition task examines the knowledge needed to recognize and select a target from an array of choices, whereas a word recall task assesses knowledge needed for target retrieval after certain cues such as a picture or the word meaning are presented. Overall, word recognition has been shown to precede the acquisition of word recall (González-Fernández & Schmitt, 2020). Using recognition and recall tasks to assess form-meaning knowledge (see Supplementary Table S1), previous studies (Aviad-Levitzky et al., 2019; Laufer & Aviad-Levitzky, 2017; Laufer & Goldstein, 2004; Schmitt, 2010) revealed that mastery levels of form-meaning knowledge are implicationally scaled, whereby meaning recognition is usually acquired before form recognition, followed by meaning recall and form recall (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004). Therefore, later-acquired form-meaning knowledge, such as recalling the meaning of a word, depends on form-meaning knowledge acquired earlier, such as the knowledge of form and meaning recognition of the same word. Nevertheless, strong correlations were found across these aspects of word knowledge (González-Fernández, 2022). A person who scores high in one aspect of word knowledge could be expected to score high in another aspect.

Following Laufer and Goldstein's (2004) framework (see Supplementary Table S1), most vocabulary tests to date (e.g., LexTALE: Lemhofer & Broersma, 2012; Vocabulary Size Test: Nation & Beglar, 2007; Updated Vocabulary Levels Test: Webb et al., 2017) assess vocabulary knowledge by measuring the number of words a test taker knows (vocabulary size) at specific mastery levels of form-meaning knowledge. Interpretation of these vocabulary test scores depends on the test format and the aspect of form-meaning knowledge being tested. For instance, the Updated Vocabulary Levels Test (Webb et al., 2017; see Nation, 1983 and Schmitt et al., 2001 for the earlier versions) was developed to assess test takers' meaning recognition at the first five 1,000-word frequency levels from the British National Corpus/Corpus of Contemporary American English (Nation, 2012). The test uses a meaning recognition matching format, in which three word meanings and six word forms (three targets and three foils) are presented together. Test takers are tasked to select the word form that matches with each of the meanings provided. The test score (matching accuracy out of 30 items) for each frequency level serves as a measure of the mastery of L2 vocabulary knowledge at specific frequency levels. Therefore, the test has been used to advise language teachers on the most appropriate word frequency level to maximize test takers' vocabulary learning. In contrast, the Vocabulary Size Test (Nation & Beglar, 2007) was designed to provide an estimate of English L1 and L2 speakers' overall receptive vocabulary size. The meaning recognition task contains 140 items that examine knowledge of English words from a wide word frequency range (1,000–14,000 frequency levels, 10 items at each 1,000 frequency level). Target words are presented in a single non-defining context one at a time, together with four meaning choices. Test takers are required to identify the meaning that matches the target word presented. Taken together, despite being designed to measure vocabulary size, the scores from different vocabulary tests can be used and interpreted differently, depending on the aspects of form-meaning knowledge and levels of language constructs (e.g., word frequency) being tested.

The Updated Vocabulary Levels Test (Webb et al., 2017) and Vocabulary Size Test (Nation & Beglar, 2007) are widely used in language classrooms because they were found to reliably predict reading ability (e.g., Laufer & Aviad-Levitzky, 2017). However, some drawbacks limit the tests' utility in a research setting. The tests take a long time to administer because test items are presented with choices in non-defining sentence context (e.g., 40 min for the Vocabulary Size Test, Nation & Beglar, 2007). Furthermore, these tests require test takers to read and understand the choices (meanings) written in sentences and match them with knowledge of the target word. As a result, the language processes involved become much more complicated and ambiguous, raising the question as to whether other language abilities (e.g., sentence comprehension or grammatical knowledge) also contribute to or affect the test scores (Meara & Miralpeix, 2016).

The quick yes/no vocabulary tests that employ an unspeeded lexical decision format (e.g., LexTALE: Lemhofer & Broersma, 2012) or a lexical judgment format (e.g., V_YesNo: Meara & Miralpeix, 2016) serve as a better alternative to assess vocabulary knowledge as a distinct construct, separated from other components of language ability. The yes/no test format was originally used as a measure of L1 vocabulary size (e.g., Anderson & Freebody, 1983) and later adopted by Meara and Jones (1988) to measure L2 vocabulary size. The Meara and Jones' original test presents words and nonwords one at a time, and test takers are required to respond "yes" or "no," to indicate whether they know the meaning of the target words. The inclusion of nonwords ensures that every test item is checked against the lexical entries in the mental lexicon, thereby requiring test takers to deploy their lexical knowledge when performing the test (Harrington, 2018). This format allows for many word items to be tested in a short amount of time and it is easy to construct and administer (Meara & Miralpeix, 2016). Furthermore, the target words are tested in a de-contextualized manner, which provides a more direct testing of the test takers' word knowledge while limiting the involvement of other language abilities (Read, 2000).

The yes/no vocabulary test format, however, despite its simplicity, is not without flaws. Yes/no vocabulary tests have been advocated as a practical measure of vocabulary knowledge because many words can be tested in a short period of time (Meara & Miralpeix, 2016). Recent studies, however, have raised some concerns about the utility of the test format. Specifically, McLean et al. (2020) have shown that a larger number of words tested in the yes/no vocabulary test format might not necessarily increase its correlation with reading proficiency when compared to other form-meaning vocabulary test formats. In their study, yes/no vocabulary test items were presented in non-defining sentences, revealing the parts of speech of the items, and no correction formula was applied to adjust the scores for guessing. Hence, it remains unclear if the findings can be generalized to the commonly used yes/no vocabulary tests, given that yes/no vocabulary tests correlated well with reading comprehension when test stimuli were presented in isolation and a correction formula was used (Siegelman et al., 2024). Therefore, further research is required to shed light on the efficacy of yes/no vocabulary tests in estimating bilinguals' lexical proficiency.

Moreover, unlike meaning recognition tests that assess knowledge of recognizing word meanings from word forms (e.g., Vocabulary Size Test: Nation & Beglar, 2007), the extent of form-meaning knowledge needed to perform in the yes/no vocabulary tests remains unclear. It is difficult to infer test takers' form-meaning vocabulary knowledge from their yes/no vocabulary test scores because there is no direct demonstration of form-meaning knowledge in the tests. Furthermore, because of the variations in test instructions across yes/no vocabulary tests (e.g., V_YesNo: Meara & Miralpeix, 2016; LexTALE: Lemhöfer & Broersma, 2012), different interpretations of the scores have been proposed. Schmitt (2010) and Zhang et al. (2020), for example, proposed that

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

correct responses in yes/no vocabulary tests require meaning recall knowledge before the identity of the letter strings can be verified. McLean et al. (2020) and Elgort (2013), on the other hand, classified the test as a form recognition test, in which test takers are required to merely identify the target word forms. Overall, despite the wide utility of the yes/no vocabulary tests, additional validation of such test format is needed to better understand the relationship between yes/no vocabulary test scores and form-meaning knowledge to justify its interpretation.

1.1. The present study

This study examined the role of form-meaning vocabulary knowledge in performing a yes/no vocabulary test by investigating the relationship between form-meaning vocabulary knowledge and yes/no vocabulary test scores and the extent to which yes/no vocabulary test scores can be predicted by different form-meaning test scores. For this purpose, the Lexical Test for Malay Speakers (LexMAL, Lee et al., 2023), a Malay yes/no vocabulary test, was presented to Malay L1 and L2 speakers. As the first freely available vocabulary test in Malay, LexMAL has been shown to provide a reliable estimate for the language proficiency of Malay L1 and L2 speakers and has high sensitivity in discriminating L1 and L2 speakers (see Lee et al., 2023, for test development and validation).

1.2. The Malay language

Being a language from the Austronesian language family, Malay is the official language of four Southeast Asian countries in the Malay Archipelago (i.e., Malaysia, Singapore, Brunei and Indonesia; Nomoto et al., 2018). It is commonly studied for cross-linguistic comparisons with English (Mazlan et al., 2024; Mohamed & Jared, 2024). Although Malay and English share the same 26 letters, Malay has a shallower orthography depth, simpler syllable structures and more transparent affixation compared to English (Yap et al., 2010). The higher vowel letter-to-phoneme ratio (see Yap et al., 2010, for a review) in Malay makes it a suitable candidate for comparisons with languages from the Romance (e.g., Spanish) and Germanic (e.g., English) families. Furthermore, Malay has a more complex morphological system than English because distinct words can be formed via rule-based affixation (Mohamed et al., 2023; Yap et al., 2010). For instance, a noun (e.g., "*peninggalan*" meaning relic) can be formed by adding a noun circumfix "*peN-…-an*" to a verb "*tinggal*" meaning "stay." In a similar way, a new verb (e.g., "*meninggalkan*" meaning leave) can be formed by adding a verb circumfix "*meN-…-an*" to the word "*tinggal.*" Because of these morphological differences, Malay words can have more syllables and a wider range in word length than English words (Lee et al., 2007).

1.3. The newly developed form-meaning vocabulary tests

Four vocabulary tests were developed to assess form-meaning knowledge of LexMAL's lexical items at various levels. These tests were specifically constructed for the purposes of this study following the item-writing protocols from previous studies (e.g., González-Fernández & Schmitt, 2020; Laufer & Aviad-Levitzky, 2017; McLean et al., 2020; Nation, 2012), to ensure their validity in assessing lexical proficiency. To understand the impact of individual word knowledge on bilinguals' performance in the yes/no vocabulary tests, the same set of words used in LexMAL was tested across the four form-meaning vocabulary tests (following González-Fernández, 2022; González-Fernández & Schmitt, 2020; McLean et al., 2020). At the test level, Malay L1 and L2 speakers' scores from the four form-meaning vocabulary tests were used as predictors for a Malay yes/no vocabulary test – the Lexical Test for Malay Speakers (LexMAL, Lee et al., 2023) – to investigate the extent to which form-meaning knowledge at each mastery level can explain their performance in the yes/no vocabulary test. At the item level, the item accuracy of each target word was compared across the vocabulary tests to evaluate the contribution of form-meaning knowledge to LexMAL accuracy. Across the vocabulary tests, we expected Malay L1 speakers to score higher than the L2 speakers (Lee et al., 2023; Rahman et al., 2018). In addition, because the yes/no vocabulary test employs a recognition task (McLean et al., 2020), we expected bilinguals' meaning and form recognition knowledge to be better predictors than meaning recall and form recall knowledge of participants' yes/no vocabulary test scores.

2. Method

2.1. Participants

One hundred and sixty bi-/multilingual Malay speakers (80 Malay L1 speakers, 70 females; 80 Malay L2 speakers, 65 females) participated in the study. All participants were students or graduates of tertiary education and had a minimum "Pass (C)" qualification for the *Bahasa Melayu* (Malay) subject in the Malaysian national high school examination (commonly known as the *Sijil Pelajaran Malaysia*). The Malay L1 speakers self-reported Malay as their L1 and dominant language, whereas all Malay L2 speakers self-reported to have acquired their L1 (Mandarin) before Malay and use Mandarin as their dominant language. Importantly, the average self-rated Malay language proficiency among the Malay L1 speakers was higher than the L2 speakers, $t(156.6) = 12.00$, $p < .001$ (see Table 2 for the summary of participants' language background). They received monetary compensation for their participation.

2.2. Instrument

The present study comprised five vocabulary tests assessing different aspects of form-meaning knowledge. The same 60 words from LexMAL were tested across these vocabulary tests. Details of each vocabulary test are described in the following subsections. A language background questionnaire adapted from the Language History

Table 2. Summary of participants' language background

	Malay L1		Malay L2	
Variable	Mean	*SD*	Mean	*SD*
Age (years)	23.21	2.74	25.30	4.93
Age of acquisition (years)				
Malay	0.46	1.32	4.83	1.41
English	4.63	2.15	3.64	2.13
Mandarin			0.40	1.15
Self-rated proficiency				
Malay	6.18	0.76	4.67	0.83
English	5.03	0.64	4.94	0.84
Mandarin			6.14	0.86

Note. Self-rated proficiency was measured on a 7-point scale (1 = *very poor*, 7 = *native-like*).

Questionnaire 3 (Li et al., 2019) was also presented, to obtain information about participants' language background and experience.

2.2.1. Target words

The 60 Malay words from LexMAL (Lee et al., 2023) consisted of 31 nouns (22 root words, nine words with "*pe-…-an*" circumfix), 17 verbs (seven root words and 10 words with "*me-…-kan*" circumfix) and 12 adjectives. These words were selected based on their distinct difficulty levels and discrimination power as evaluated by item response theory analysis (see Lee et al., 2023 for LexMAL item assessment and selection). As the test was designed to assess highly proficient and moderately proficient Malay speakers, these words were a combination of high-frequency words that were most likely to be known by most speakers, as well as low-frequency words that were likely to be known only by highly proficient Malay speakers. The distribution of word stimuli across five frequency bands in Zipf values[1] (van Heuven et al., 2014) is summarized in Supplementary Table S2.

To reduce potential learning effects from repeated exposure, only half of the 60 target words from LexMAL were presented for each of the subsequent vocabulary tests. This method presented the target words only twice across all form-meaning vocabulary tests, lowering the chances of participants answering based on their memory of test items or cues from previous presentations. Two wordlists (A and B) with matched word frequency and length ($ts \leq 0.50$, $ps \geq .62$) were created from the 60 target words (see Supplementary Table S3 for lexical information of each wordlist). The presentation of wordlists was counterbalanced among the participants. They saw the same wordlist (either wordlist A or B) for the Form Recall and Form Recognition tests, and the other wordlist for the Meaning Recall and Meaning Recognition tests.[2] Thus, participants ($n = 40$ from each language group) who took the Form Recall and Form Recognition tests with wordlist A took the Meaning Recall and Meaning Recognition tests with wordlist B. In addition, another 40 Malay words that spread across the frequency bands were also selected from Yap et al. (2010) as filler items. The filler items served as distractors to further minimize testing effects from preceding tests that might arise from participants focusing solely on the target words. Each vocabulary test (except LexMAL) presented 10 novel filler items in addition to the target words from wordlist A or B. Each filler item was presented only once throughout the study. The target words and filler items were matched in terms of word frequency (Zipf value) and word length, $ts \leq 0.01$, $ps \geq .93$.

2.2.2. Vocabulary test 1: LexMAL

LexMAL (Lee et al., 2023) is an unspeeded yes/no vocabulary test designed to estimate the Malay proficiency of L1 and L2 speakers. It contains a total of 90 items (60 words and 30 nonwords). The nonwords were generated by randomly substituting one letter of Malay real words using Malay bigrams and trigrams extracted from a large Malay word list (see Lee et al., 2023, for a detailed description). Participants were required to indicate if letter strings are existing Malay words by responding "yes" or "no."

Scoring. LexMAL score (normalized Ghent score, see (1); Wen et al., 2023) was computed by summing up the number of correctly identified word stimuli and penalizes the score based on guessing by the participant ("yes" responses for nonword stimuli, i.e., false alarms). Normalized Ghent score ranges from −100% to 100%, with a negative score indicating a higher false-alarm rate than correct word identification.

$$\text{Normalized Ghent score} = \left(N_{\text{yes to word stimuli}} - 2N_{\text{yes to nonword stimuli}}\right) \times \frac{100}{60} \quad (1)$$

2.2.3. Vocabulary test 2: Form Recall

A Form Recall test was developed to assess the ability to recall the target word form from its definition. The definitions were adapted from the dominant meaning of target words provided in the Malay dictionary – *Kamus Dwibahasa* (Dewan Bahasa dan Pustaka; Ibrahim, 2002). As the test focuses on vocabulary knowledge, the definitions were rewritten in much easier language than the ones provided by the dictionary to minimize the demands on vocabulary knowledge beyond the target word (Nation, 2012). For this purpose, words from the same frequency band,[3] if not higher than the target words, were used as much as possible. When lower-frequency word types were required to describe a concept, we sought for more commonly known words (judged by word family) as far as possible. For example, the lower frequency word "*dimasak*/cooked" (Zipf value = 2.18) was used to rewrite the meaning of "*mentah*/raw" (Zipf value = 2.71), as in "*belum* ***dimasak*** *penuh*/uncooked," because its root word "*masak*/cook" (Zipf value = 3.91) is a commonly used Malay word and has a higher word frequency than "*mentah*." Two Malay L1 speakers with a background in linguistics were recruited to proofread the definitions to ensure their accuracy and that the words used in the definitions were not more difficult than the target words.

The definitions were presented one at a time and participants were required to type the target word form that corresponded to the definition provided. To avoid correct responses that were not the target words, the number of letters and the third letter of the root words were specified for each trial item. This approach was similar to González-Fernández and Schmitt (2020), Laufer and Goldstein (2004) and McLean et al. (2020). An example of the Form Recall item is given below.

bahagian badan di sebalik dada (*8 huruf*)/(body part behind the chest (eight letters)) _ _ *l* _ _ _ _ _/(back)

A pilot study was conducted to assess if the presentation of cues (number of letters and third letter of root word) would lead to ceiling performance with L1 speakers. The pilot involved eight Malay L1 speakers. Overall, participants performed significantly better when they were presented with two cues, that is, with the number of letters and the third letter shown, $M = 27.92\%$, $SD = 30.91\%$, than when they were presented with only one cue, that is, a number of letters only, $M = 15.42\%$, $SD = 25.25\%$, $t(113.48) = 2.43$, $p = 0.02$. As the mean accuracy for the two-cues group was still far lower than a ceiling performance, both cues were presented together with the definitions in the

[1] In the present study, we report and run analyses of word frequency in Zipf values because the Zipf scale offers a more transparent and intuitive interpretation of word frequency (Brysbaert et al., 2018; van Heuven et al., 2014). Zipf values vary between 1 (0.01 frequency per million) and 7 (10,000 frequency per million). Low-frequency word have Zipf values of 3 or lower and high-frequency words have Zipf values of 4 or higher (see van Heuven et al., 2014).

[2] To distinguish between the form-meaning tests devised for this study and the form-meaning knowledge assessed as a latent construct, the form-meaning tests are capitalized whenever we refer to the tests.

[3] Due to the limited number of words covered in Yap et al. (2010), we also referred to the DBP Corpus Database (Rusli et al., 2006) for word frequency information for some uncovered words during this screening procedure.

Form Recall test, to ensure that the test is not too difficult for the L2 speakers.

Scoring. The responses were scored dichotomously and only answers that matched the target words and were spelt correctly were marked as correct. The percentage of correct responses was used to compute the Form Recall score.

2.2.4. Vocabulary test 3: Meaning Recall

The Meaning Recall test is an open-ended written test, in which the ability to recall the meaning of the target word based on its word form was assessed. The target word forms were presented one at a time (e.g., "*canggih*" meaning sophisticated) and participants were required to type the meaning of the target word in any language they know (i.e., Malay, English or Mandarin), either in the form of a translation, a synonym, a description, a definition or a sentence, as long as the specific meaning tested was clearly demonstrated (following González-Fernández & Schmitt, 2020; Laufer & Aviad-Levitzky, 2017; McLean et al., 2020).

Scoring. The responses were scored dichotomously. Responses were scored as correct if participants provided a correct synonym, translation or description of that meaning. For example, if a participant supplied "*paragraf*" as a synonym to "*perenggan*," "paragraph" as a translation, or described "*perenggan*/paragraph" as "*bahagian penulisan yang mengandungi beberapa baris ayat*/a piece of writing with several sentences" in any of the three languages, the response was scored as correct. Conversely, translations or descriptions that were too general or did not reflect the meaning of the target word were considered incorrect (e.g., providing "passage" or "*berkaitan dengan karangan*/related to essay" for the target word "*perenggan*/paragraph"). To ensure scoring reliability, a proficient Malay L1 (LexMAL score = 90.0%) and a Mandarin L1 speaker who also speaks Malay as L2 (LexMAL score = 50.0%)[4] with a linguistics background were trained and scored responses from a random 20% of speakers selected from each respective language group (n = 16 each). All responses from the selected participants were scored (n = 40 each) and only the responses accepted by both the scorer and corresponding author were considered correct. Overall, the L1 responses were scored with 97.2% agreement and a Cohen's kappa of 0.94, whereas L2 responses were scored with 91.2% agreement and a Cohen's kappa of 0.81.

2.2.5. Vocabulary test 4: Form Recognition

The Form Recognition test assesses the ability to recognize a Malay word form given its meaning in Malay. This test adopted a multiple-choice format, where participants were presented with the same definitions they saw in the Form Recall test (except for filler items) and were asked to choose the target word form that matched each definition. The target words were presented with three foils. In accordance with Nation (2012) and McLean et al. (2020), the foils presented were of the same frequency band and word class as the target word. Words that shared core elements of meaning with the target word were avoided to account for partial knowledge by avoiding confusion caused by words with related meaning (Nation, 2012). For example, the item testing "*gerbang*/archway" did not include foils that require participants to distinguish between various types of doors or gates. The two Malay L1 speakers who reviewed the Form Recall test also reviewed the foils, to ensure that there was no other possible answer among the foils other than the target word form. An example of a Form Recognition item is presented below.

gambaran tentang masa depan yang terbayang dalam fikiran /(an image of the future that appears in the mind)

A. *leret*/(swipe)
B. *nyawa*/(life)
C. ***angan*/(wish)**
D. *tongkah*/(stick)

Scoring. The responses were scored dichotomously and the percentage of correct responses was used to compute the Form Recognition score.

2.2.6. Vocabulary test 5: Meaning Recognition

The Meaning Recognition test assesses the ability to identify the meaning of a target word form from a list of four choices. The same foils selected for the Form Recognition test were used in this test and their meanings were presented as the other three possible answers for each target word form. Meanings of the target words and foils were written using the same criteria as described for the Form Recall test. In accordance with Nation (2012), non-meaning clues such as the length of the choice and general versus specific choices were avoided when writing the definitions. This was later confirmed by the two Malay L1 speakers who reviewed the definitions. An example of Meaning Recognition is given below.

bahang/(heat)

A. ***rasa panas dari benda hangat*/(hot sensation from warm objects)**
B. *benda-benda yang dibuang*/(discarded objects)
C. *bayaran perjalanan*/(travel fees)
D. *harapan supaya sesuatu menjadi*/(hope that something will happen)

Scoring. The responses were scored dichotomously, and the percentage of correct responses was used to compute the Meaning Recognition score.

2.2.7. Language background questionnaire

We selected the most relevant questions from the Language History Questionnaire 3 (Li et al., 2019), which focused on information about participants' multilingual language history and experience, such as participants' age of acquisition, education history, and years and context of learning experience for all their known languages. The questionnaire also asked for self-rated reading, writing, listening and speaking proficiency in Malay, English and Mandarin (Mandarin L1 participants only), using a scale of 1 (*very poor*) to 7 (*native-like*).

2.3. General procedure

The present study was administered fully online using Qualtrics (https://www.qualtrics.com). Participants were instructed to complete all tasks without external aids (e.g., dictionary) and they were given as much time as needed to complete the study. The study was approved by the Ethics Committee in the School of Psychology at the University of Nottingham Malaysia (Application Identification Number: LST220222). Written

[4] The Malay L2 scorer also used Mandarin as L1. LexMAL score of 50% is well above the mean score of the L2 speakers in the present study (see Table 3). This scorer was tasked to score the Mandarin/Malay/English responses collected from the Malay L2 speakers.

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

consent was acquired from participants before data collection started.

The presentation order of the vocabulary tests was based on the difficulty hierarchy of form-meaning knowledge, progressing from the most difficult to the easiest (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; McLean et al., 2020). This approach ensured that word exposure in the earlier vocabulary tests did not affect participants' responses in the later tests (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; McLean et al., 2020; Nation, 2013; Nation & Webb, 2011; Schmitt, 2010). The study started with LexMAL, in which the participants were required to make yes/no decisions to every stimulus presented to them, one at a time. The words and nonwords were presented to all participants in the same randomized order. Participants were required to indicate "yes" if they thought the letter string presented on the screen was an existing Malay word. They were told to respond "yes" to the stimulus even if they did not know the exact meaning of the letter string but were certain that it was an existing Malay word. In cases where they thought the letter string was not a Malay word, or they were in doubt, they were instructed to respond "no." They were also reminded that errors were penalized to control for response bias. At this point of testing, information about form-meaning links was not revealed to the participants. No feedback was provided to the participants so that the unknown words remained unknown to them.

After LexMAL, a non-language filler task with 10 items adapted from Raven's progressive matrices task (Raven, 2000) was presented. This task presented shapes in a 3 × 3 matrix with a blank on the lower right field, in which participants are required to deduct the rules of the matrix and select the shape that best fits the blank from an array of choices. Following the filler task, the other four vocabulary tests were presented according to the hierarchy of difficulty of form-meaning knowledge (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; McLean et al., 2020). The testing started with the Form Recall test, followed by the Meaning Recall test, Form Recognition test and Meaning Recognition test. By moving down the theoretical hierarchy of difficulty, it was unlikely for a previous test to inform the subsequent test. Participants were presented with stimuli from different wordlists across vocabulary tests (e.g., participants who saw the definitions from wordlist A in the form recall test were tested on the production of meaning of target words from wordlist B) and the stimuli presentation order was randomized.

Each vocabulary test started with specific instructions on how to complete it and examples illustrating how to respond to the items. Instructions were presented in Malay for all the vocabulary tests. Participants were unable to go back to a previous item once they submitted an answer to avoid cross-contamination of responses between vocabulary tests and items within a test. After the vocabulary tests, participants completed the language background questionnaire as the last part of the study.

3. Results

The mean total duration[5] for the participants to complete LexMAL, Form Recall, Meaning Recall, Form Recognition and Meaning Recognition tests were 5.31 (SD = 3.00), 36.17 (SD = 17.28), 14.92 (SD = 10.18), 5.26 (SD = 2.39) and 6.05 (SD = 2.72) min, respectively. Participants' mean test scores are summarized in Table 3. Before applying the correction formula, both Malay L1 and L2 speakers obtained relatively high raw scores (hit rate or the percentage of correctly identified LexMAL word items). However, false alarm rates were also high for both groups, reflecting a considerable amount of guessing. Consequently, relying on uncorrected raw scores could overestimate participants' vocabulary knowledge, as raw scores might simply be elevated by a tendency to respond "yes" frequently, as evidenced by the high number of "yes" responses to nonwords. To address this, the normalized Ghent score was computed by adjusting raw scores by accounting for the false alarm rates or response bias (tendency to respond "yes" to test items). Steiger's (1980) z-tests for dependent correlations, conducted using the *cocor* R package (Diedenhofen & Musch, 2015), confirmed that the normalized Ghent score had a significantly stronger correlation with both form and meaning recognition test scores compared to the raw score correlations with these tests, $zs \geq 2.90$, $ps \leq .004$. Therefore, the normalized Ghent score was used for the subsequent analyses.

Overall, L1 speakers appeared to score higher than L2 speakers across all vocabulary tests and the test scores for LexMAL, Meaning Recognition and Form Recognition appeared higher than Meaning Recall and Form Recall (see Figure 1 for the boxplot). A fixed-effects hierarchical regression analysis was conducted to examine if the four vocabulary test scores predict LexMAL accuracy. Subsequently, a generalized mixed-effects model was conducted to assess if form-meaning knowledge demonstrated in each vocabulary test could predict LexMAL item accuracy and at the same time investigate language dominance effect across the vocabulary tests. Lastly, the receiver operating characteristic (ROC) curve analyses (Lalkhen & McCluskey, 2008; Read et al., 2015) were conducted to examine if the vocabulary tests were able to discriminate between the vocabulary knowledge of Malay L1 and L2 speakers. The internal reliability for all tests was computed using Cronbach's alpha. All vocabulary tests had Cronbach's alpha >.80, indicating good internal reliability (see Supplementary Table S6).

3.1. Predictive power of vocabulary knowledge on LexMAL

Correlation analysis conducted using R (version 4.1.1; R Core Team, 2021) revealed that the scores of form-meaning vocabulary tests and LexMAL were positively correlated. All the correlations were significant (in all cases $p < .001$; see Figure 2 for the correlation matrix). To compare the strength of correlations between LexMAL scores and each form-meaning vocabulary test, Steiger's (1980) z-tests for dependent correlations were conducted using the *cocor* R package (Diedenhofen & Musch, 2015). These findings suggest that LexMAL scores are more strongly correlated with recognition knowledge than with recall knowledge. There was no significant difference in the correlation strengths between LexMAL scores and the two form and meaning recognition test scores, nor did the correlation strengths differ between LexMAL scores and the two form and meaning recall test scores, $zs \leq 0.03$, $ps \geq .97$.

Linear regression analyses were conducted to examine how well LexMAL scores can predict Meaning Recognition, Form

[5] The response time data were collected in an online study. Participants were given as much time as they needed to answer the vocabulary tests, which could result in less reliable response time. Therefore, caution should be taken when interpreting them as direct indicators of time needed to test a specific amount of word items.

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

Table 3. Means and *SD*s (in percentage) of accuracy for each vocabulary test

	Malay L1 (*N* = 80)			Malay L2 (*N* = 80)		
Vocabulary test	*M*	*SD*	Range	*M*	*SD*	Range
LexMAL	74.12	19.45	11.67–100.00	34.15	21.78	0.00–96.67
Hit rate	88.83	7.61	66.67–100.00	68.69	16.38	33.33–100.00
False alarm rate	14.71	18.66	0.00–76.67	34.54	20.67	0.00–86.67
Form Recall	38.22	13.87	0.00–70.00	23.19	15.59	0.00–72.50
Meaning Recall	47.53	14.19	10.00–85.00	33.22	17.56	7.50–85.00
Form Recognition	92.69	5.29	72.50–100.00	74.44	14.14	32.50–100.00
Meaning Recognition	88.34	7.58	47.50–100.00	62.53	17.25	20.00–100.00

Note. Hit rate is the percentage of correctly identified word items in LexMAL. False alarm rate is the percentage of incorrectly identified nonword items in LexMAL.

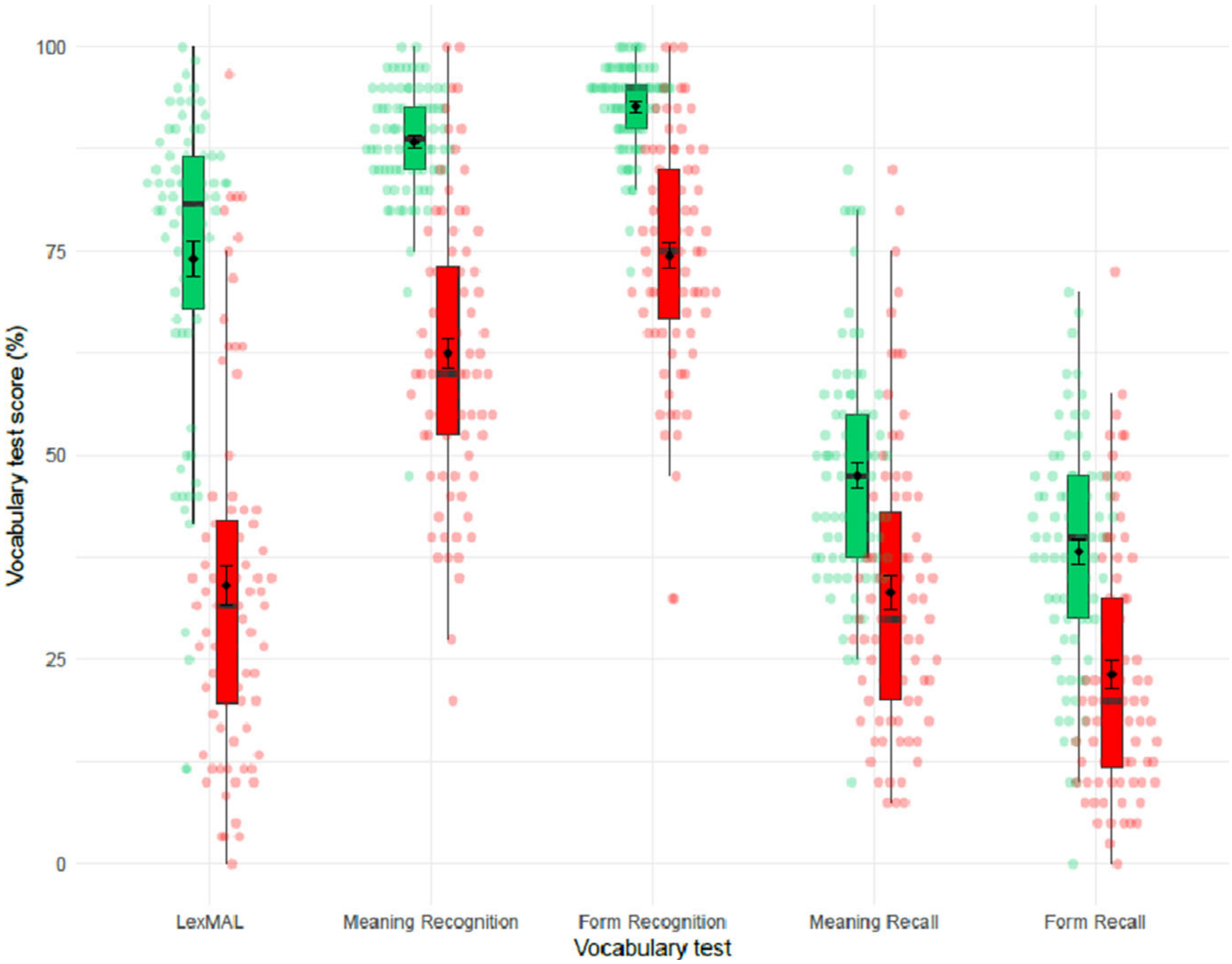

Figure 1. Vocabulary test scores of the two language groups. *Note*: *Green* represents the L1 speakers and *red* represents the L2 speakers. *Black* dots denote the group means, with SEs denoted by the whiskers.

Recognition, Meaning Recall and Form Recall scores. The findings revealed that LexMAL scores significantly predicted the scores of Meaning Recognition (R^2 = .53, F(1, 158) = 174.70, p < .001; adjusted R^2 = .52), Form Recognition (R^2 = .53, F(1, 158) = 175.73, p < .001; adjusted R^2 = .52), Meaning Recall (R^2 = .34, F(1, 158) = 79.88, p < .001; adjusted R^2 = .33) and Form Recall (R^2 = .34, F(1, 158) = 80.37, p < .001; adjusted R^2 = .33). Of all the models, the Form Recognition model had the lowest Akaike Information Criterion and Bayesian Information Criterion values, indicating it is the best-fitting model. This suggests that LexMAL scores are particularly effective at predicting Form Recognition performance.

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

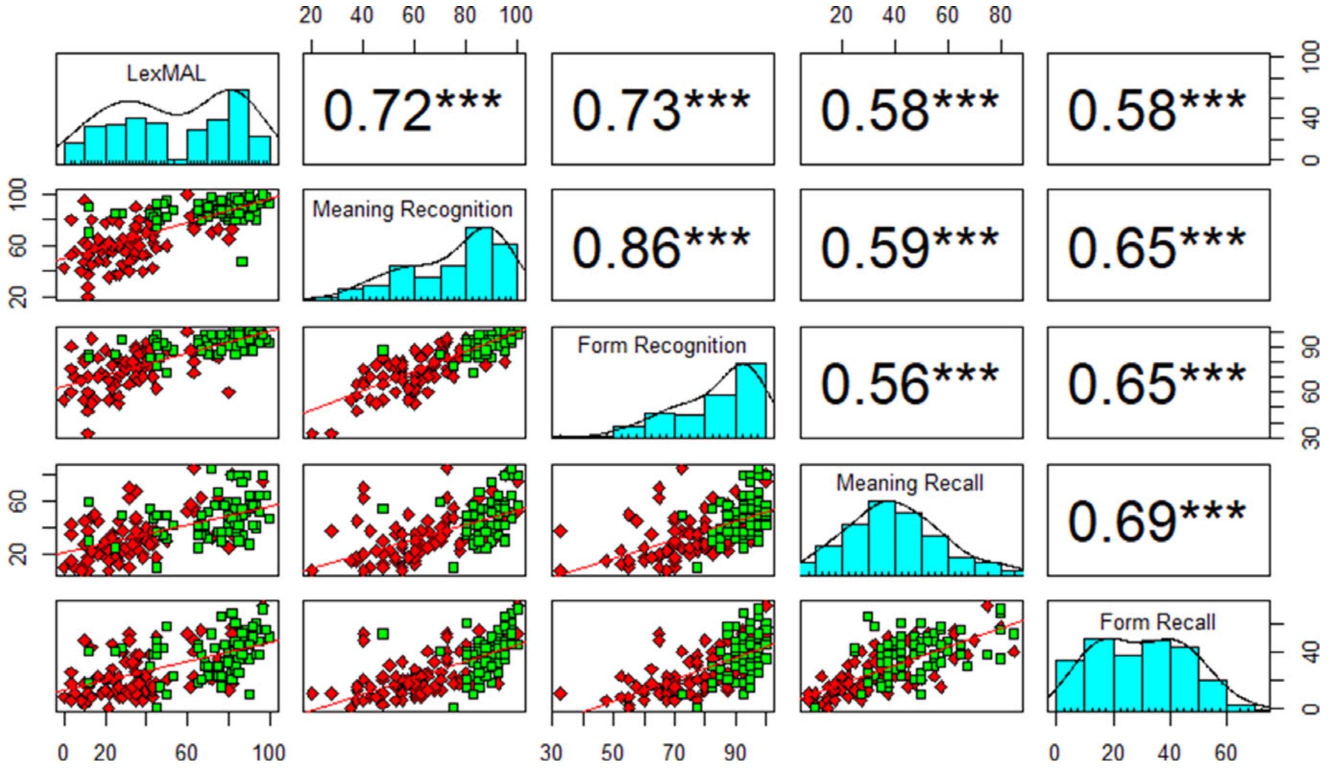

Figure 2. Correlation of scores between LexMAL and form-meaning vocabulary tests. *Note*: *Green* points represent the L1 speakers and *red* points represent the L2 speakers.

However, it is important to note that form-meaning knowledge is highly correlated (*rs* ≥ .56; see Figure 2). This correlation suggests that the explanatory power observed in the separate regression models is likely to be affected by the vocabulary knowledge aspects shared across the form-meaning vocabulary tests. To assess if different aspects of form-meaning knowledge can account for a significant proportion of the variance in LexMAL score, fixed-effects hierarchical regression analysis was conducted with LexMAL score as the dependent variable and test scores from different aspects of form-meaning knowledge as fixed effects. The aspects of form-meaning knowledge were entered one by one into the model according to the acquisition order (Aviad-Levitzky et al., 2019; Laufer & Aviad-Levitzky, 2017; Laufer & Goldstein, 2004; Schmitt, 2010). Meaning Recognition score was entered in the first step to predict the LexMAL score, followed by Form Recognition, Meaning Recall and Form Recall scores in the second, third and fourth steps, respectively.

The first three regression models, at each step, explained significantly more variance than its preceding model(s), *Fs* ≥ 11.89, *ps* < .001. The Form Recall score added to the final step did not account for additional variance in the LexMAL score, *F* = 0.11, *p* = .74 (see Supplementary Table S4 for the model statistics). The third model was the best-fit model explaining 59% of the variance in LexMAL score, *F*(3, 156) = 75.96, *p* < .001, Cohen's f^2 = 1.44. The semi-partial correlation squared for Meaning Recognition, Form Recognition and Meaning Recall scores were 27.94%, 39.68%, and 32.38% respectively.

3.2. Predictive power of language dominance and vocabulary knowledge on item accuracy

To investigate if language dominance (L1 or L2) and form-meaning knowledge of the target words at various knowledge aspects (measured by form-meaning vocabulary tests) predict item accuracy, generalized mixed-effects modeling was conducted using the *lme4* R package (Bates et al., 2015). The fixed effects in the model were the language dominance group (deviation coding of 0.5 for L1 speakers and −0.5 for L2 speakers) and vocabulary tests (deviation coding of 0.8 for the target vocabulary test and −0.2 for the non-target vocabulary tests) as well as the interaction between these predictors. LexMAL was set as the baseline of comparison for the vocabulary tests. The model was fitted with participants and stimuli as random effects. As the scores from different vocabulary tests were highly correlated (the highest correlation was between Meaning Recognition and Form Recognition, *r* = .86; see Figure 2) and could induce collinearity issue, random intercepts and slopes were fitted with no correlation[6] (zero-correlation parameter for random effects). Within-subject predictors (i.e., the vocabulary tests) were included as by-subject random slopes, and language dominance group, vocabulary tests, as well as their interaction were included as by-item random slopes.

The generalized mixed-effects model revealed that language dominance affected vocabulary test accuracy ($\beta = 1.70$, $SE = 0.18$, $z = 9.69$, $p < .001$). For the same test items that were correctly identified in LexMAL, L1 speakers had a higher tendency than L2 speakers to correctly answer these items in the form-meaning vocabulary tests.

The main effects of vocabulary tests were also indicated. When test items were correctly identified in LexMAL, their log odds of being correctly answered in other vocabulary tests were higher in the Form Recognition test ($\beta = .66$, $SE = 0.18$, $z = 3.58$,

[6] This random-effect structure helps to answer our research question if each of the vocabulary tests could predict item accuracy in LexMAL, instead of its unique contribution to predict LexMAL accuracy while taking other vocabulary tests into consideration.

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

$p < .001$) but lower in the Meaning Recall ($\beta = -2.88$, $SE = 0.17$, $z = -17.40$, $p < .001$) and Form Recall tests ($\beta = -3.69$, $SE = 0.22$, $z = -16.47$, $p < .001$). The log odds for Meaning Recognition were not significant ($\beta = -.20$, $SE = 0.14$, $z = -1.43$, $p = .15$), suggesting that there was no clear indication that correct identifications of real words in LexMAL would predict their meaning being recognized in the Meaning Recognition test. Furthermore, significant interaction effects between the language group and target test were found (see Figure 3) in the Meaning Recall ($\beta = -.99$, $SE = 0.23$, $z = -4.37$, $p < .001$) and Form Recall tests ($\beta = -.63$, $SE = 0.21$, $z = -2.96$, $p = .003$). The target test factor compares the odds ratio of the target and non-target vocabulary tests. Importantly, independent of the interactions, vocabulary items correctly identified in LexMAL were likely to be correctly answered in the Form Recognition test regardless of the language group. Table 4 provides an overview of the estimates of fixed effects and the interactions.

Using *emmeans* R package (Lenth, 2023), post-hoc pairwise comparisons were conducted to examine how language groups interacted with the target tests (i.e., Meaning Recall and Form Recall; see Table 5 for test statistics). In summary, the L2 speakers were less likely than the L1 speakers to score the correctly identified LexMAL items in both levels (target and non-target vocabulary tests) of Meaning Recall and Form Recall tests, $ps < .01$, corrected with Tukey adjustment. Within each language group, participants were more likely to score in the non-target vocabulary tests in comparison to the Meaning Recall and Form Recall tests, $ps < .001$, indicating their poorer performance with the Meaning Recall and Form Recall tests. Specifically, the likelihood of scoring in the non-target tests was at least 29.35 times higher than in the target tests for the L1 speakers, indicating that the effects of target tests were stronger for L1 speakers than for L2 speakers (whose target and non-target tests' odds ratio was 29.18 at highest; see odds ratio in Table 5). The estimated marginal means and *SE*s for each pairwise combination are summarized in Supplementary Table S5.

3.3. Discriminant ability and reliability of vocabulary tests

To examine if the vocabulary tests can distinguish between L1 and L2 speakers' vocabulary knowledge, ROC curve analyses were conducted using the *pROC* R package (Robin et al., 2011). ROC curve analysis is commonly used in clinical testing to assess the accuracy of a diagnostic test in diagnosing clinical disorders (Lalkhen & McCluskey, 2008; Read et al., 2015). The ability of a test to discriminate between people with and without a disorder, or the discrimination power, is calculated using the area under the curve (AUC). The ROC curve plots the true positive rate or test's

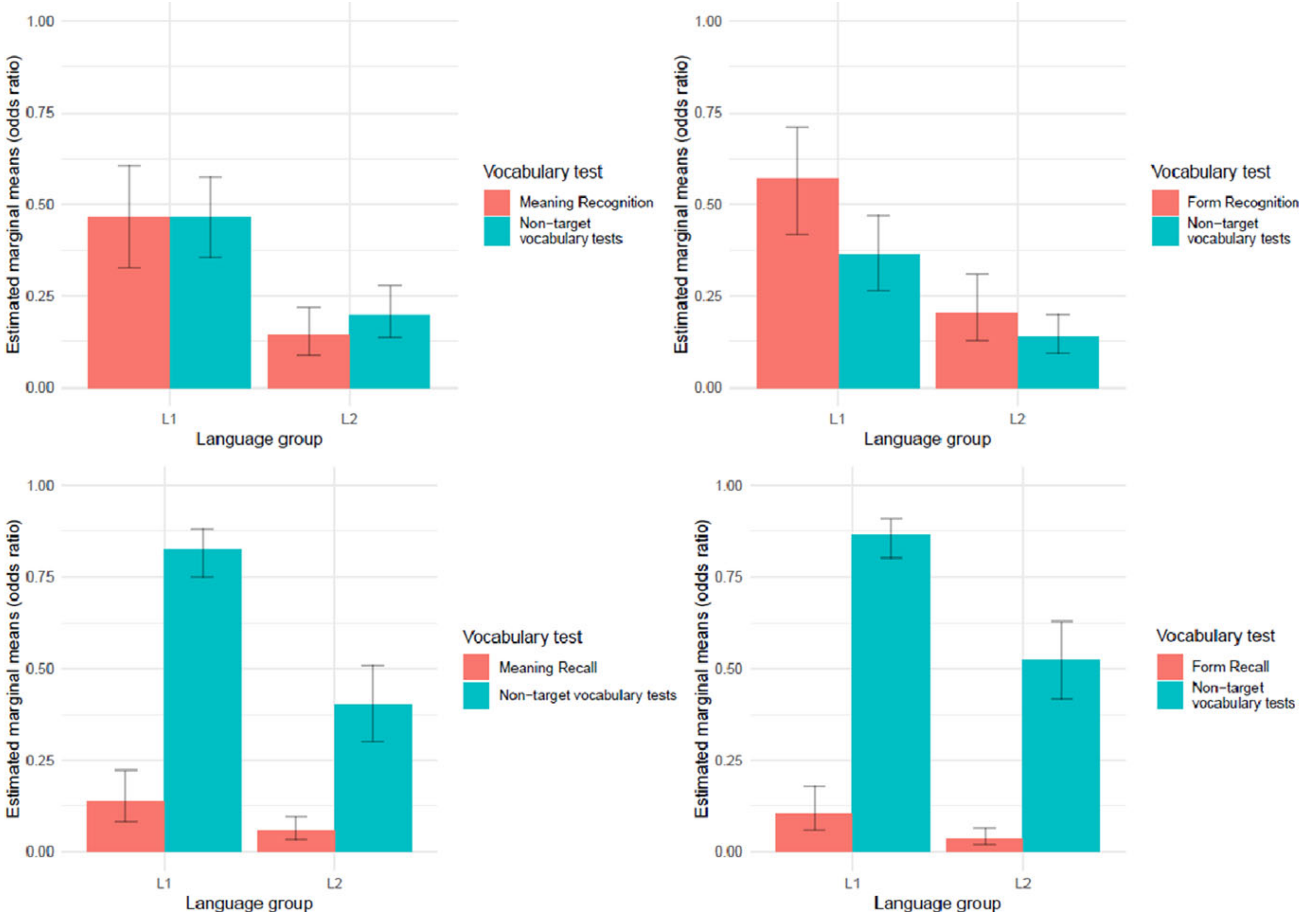

Figure 3. Marginal effects of two-way interaction between language group and odds ratio of item accuracy. *Note*: Language group and vocabulary test were contrast coded; 0.5 for L1 speakers and −0.5 for L2 speakers; 0.8 for the target vocabulary test and −0.2 for the non-target vocabulary tests. For example, in the bottom-left plot, Form Recall is the target vocabulary test, whereas Meaning Recognition, Form Recognition and Meaning Recall are the non-target vocabulary tests. The odds of correctly scoring the vocabulary items correctly identified in LexMAL was lower in Form Recall to the average odds ratio of the non-target vocabulary tests across language groups. Particularly, the difference in odds ratio was greater in L1 than L2 group.

Table 4. Summary of the generalized mixed-effects model

	Item accuracy		
Predictors	Odds ratio	CI	*p*
(Intercept)	2.62	1.88–3.66	**<.001**
Language group	5.47	3.88–7.71	**<.001**
Meaning Recognition versus LexMAL	0.82	0.62–1.08	.153
Form Recognition versus LexMAL	1.94	1.35–2.78	**<.001**
Meaning Recall versus LexMAL	0.06	0.04–0.08	**<.001**
Form Recall versus LexMAL	0.03	0.02–0.04	**<.001**
Language group × Meaning Recognition	1.50	0.97–2.31	.066
Language group × Form Recognition	1.47	0.91–2.36	.114
Language group × Meaning Recall	0.37	0.24–0.58	**<.001**
Language group × Form Recall	0.53	0.35–0.81	**.003**
Random effects			
σ^2		3.29	
$\tau_{00\ \text{participant}}$		0.67	
$\tau_{00\ \text{stimuli}}$		1.45	
$\tau_{11\ \text{participant.meaningrecognition}}$		0.66	
$\tau_{11\ \text{participant.formrecognition}}$		0.63	
$\tau_{11\ \text{participant.meaningrecall}}$		0.63	
$\tau_{11\ \text{participant.formrecall}}$		0.56	
$\tau_{11\ \text{stimuli.languagegroup}}$		0.70	
$\tau_{11\ \text{stimuli.meaningrecognition}}$		0.64	
$\tau_{11\ \text{stimuli.formrecognition}}$		1.38	
$\tau_{11\ \text{stimuli.meaningrecall}}$		1.20	
$\tau_{11\ \text{stimuli.formrecall}}$		2.49	
$\tau_{11\ \text{stimuli.languagegroup:meaningrecognition}}$		0.84	
$\tau_{11\ \text{stimuli.languagegroup:formrecognition}}$		1.10	
$\tau_{11\ \text{stimuli.languagegroup:meaningrecall}}$		1.40	
$\tau_{11\ \text{stimuli.languagegroup:formrecall}}$		0.86	
ICC		0.39	
$N_{\text{participant}}$		160	
N_{stimuli}		60	
Observations		28800	
Marginal R^2/Conditional R^2		0.393/0.631	

Note. σ^2 = residual error, τ_{00} = variance of random intercepts, τ_{11} = variance of random slopes. LexMAL was the baseline for vocabulary test comparison.
The bold values indicate statistical significance, i.e., p <.05.

sensitivity (e.g., the accuracy of a test in identifying people with a disorder) on the *y*-axis and the false positive rate (1 – specificity, with specificity referring to the accuracy of a test in identifying people without a disorder) on the *x*-axis.

ROC curve analyses for the vocabulary tests (see Figure 4) revealed that LexMAL and Meaning Recognition tests had a very good ability to discriminate vocabulary knowledge of Malay L1 and L2 speakers, as indicated by its AUCs > 0.90. The Recognition test's discriminant ability was good with an AUC of 0.89. Meaning Recall and Form Recall tests, on the other hand, had fair discriminant ability with AUCs > 0.75. The ROC curve analyses identify the optimal cutoff score for each vocabulary test that yields the highest test sensitivity (the test's ability to correctly identify L1 speakers) and specificity (the test's ability to correctly identify L2 speakers) (see Figure 4, right panel). Of all the vocabulary tests, the sensitivity and specificity of LexMAL, Meaning Recognition and Form Recognition were higher than 80%, with the cutoff scores being set at 64.17%, 81.25%, and 88.75%, respectively. This finding suggests that LexMAL and the recognition tests were able to identify L1 and L2 speakers at least 80% correctly using the cutoff scores. Importantly, the sensitivity of the recall tests was lower than 70%, indicating that the tests were less accurate in identifying L1 speakers compared to the LexMAL and recognition tests.

Table 5. Summary of test statistics for pairwise comparisons between language group, Meaning Recall and Form Recall

Comparison group	Odds ratio	*SE*	*z*
Meaning Recall			
L2-T/L1-T	0.39	0.12	-3.15^{**}
L2-NT/L1-NT	0.14	0.03	-8.72^{***}
L1-NT/L1-T	29.35	5.99	16.55^{***}
L2-NT/L2-T	10.86	2.15	12.06^{***}
Form Recall			
L2-T/L1-T	0.32	0.10	-3.83^{***}
L2-NT/L1-NT	0.17	0.04	-7.91^{***}
L1-NT/L1-T	54.57	13.61	16.04^{***}
L2-NT/L2-T	29.18	7.17	13.72^{***}

Note. T = target vocabulary test; NT = non-target vocabulary tests. Non-target vocabulary tests include all form-meaning vocabulary tests except the target vocabulary test.
$^{**}p < .01$.
$^{***}p < .001$.

4. Discussion

The present study used four vocabulary tests to examine the contribution of bilinguals' form-meaning knowledge to their accuracy in a yes/no vocabulary test. In addition to significant correlations between the vocabulary test scores, our findings revealed that all form-meaning vocabulary test scores (except Form Recall) predicted yes/no vocabulary test scores. All the vocabulary tests were shown to have good discriminant ability between L1 and L2 speakers, AUCs > 0.75. Importantly, bilinguals' form-meaning knowledge, specifically form recognition, meaning recall, and form recall, were shown to predict bilinguals' item accuracy across the vocabulary tests.

At the test level, the best-fit fixed-effects hierarchical regression model showed that test scores from Meaning Recognition, Form Recognition and Meaning Recall accounted for 59% of the variance in LexMAL scores. In addition, the semi-partial correlation squared revealed Form Recognition accuracy as the strongest unique predictor, followed by Meaning Recognition and Meaning Recall accuracy. This corroborates with existing literature that both form and meaning knowledge has a unique contribution to lexical proficiency (e.g., González-Fernández, 2022; González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; Nation, 2013, 2020). Meaning Recognition and Form Recognition tests, despite having a high correlation between the test scores, measure distinct aspects of vocabulary knowledge (González-Fernández & Schmitt, 2020). Furthermore, meaning recall but not form recall explained a

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

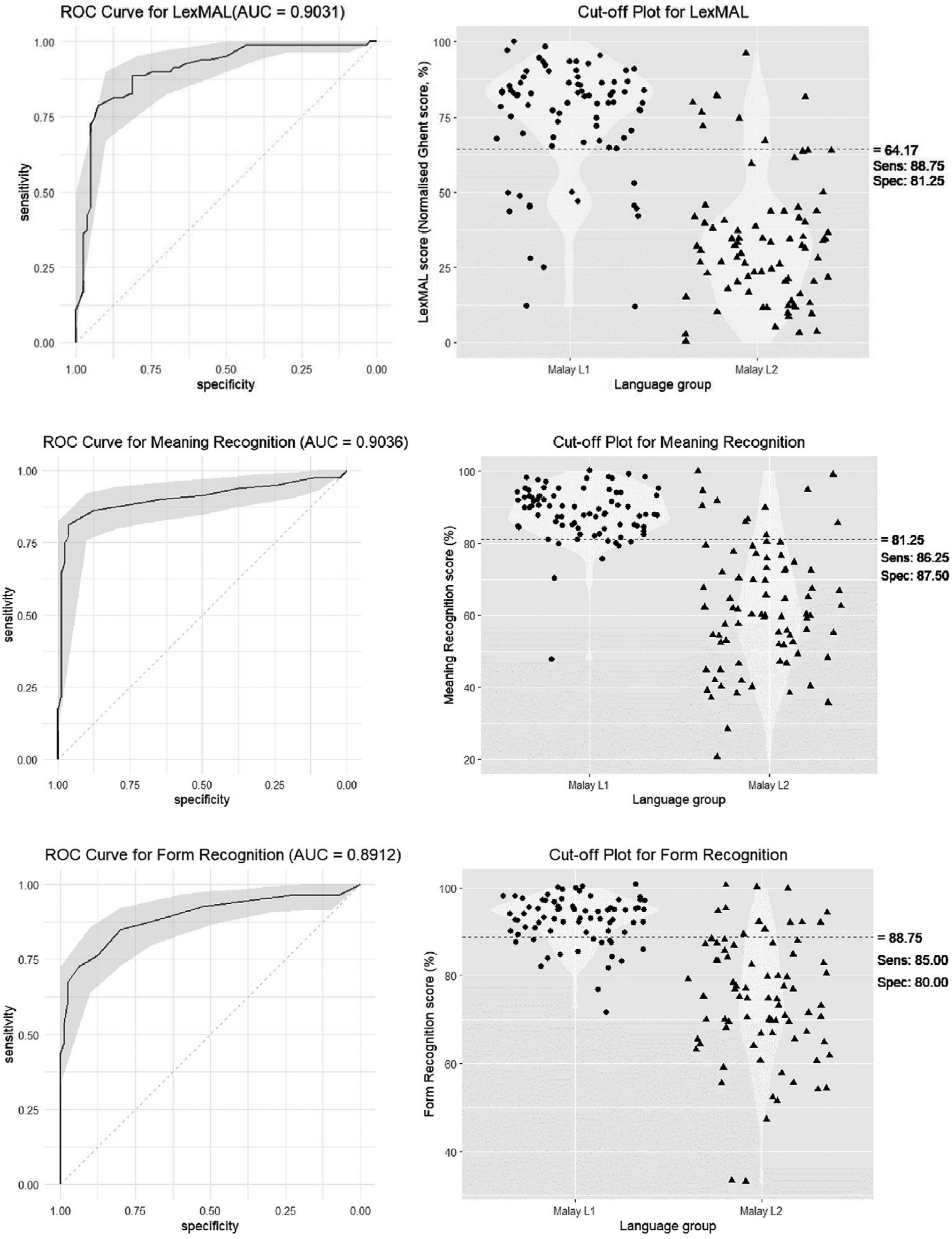

Figure 4. ROC curve for the vocabulary tests. *Note*: The left panel shows the ROC curve for the LexMAL vocabulary test, plotting sensitivity (true positive rate) against 1 – specificity (false positive rate). AUC represents the discriminatory power of tests. For example, the LexMAL test has an AUC of 0.9031, indicating that LexMAL scores correctly discriminate between Malay L1 and L2 speakers 90.31% of the time. The right panel presents the distribution of test scores for each vocabulary test. The dashed horizontal line represents the optimal cutoff score for distinguishing between Malay L1 and L2 speakers. Sensitivity represents the accuracy of the test in identifying L1 speakers, while specificity indicates the accuracy of the test in identifying L2 speakers. For instance, for LexMAL, a cutoff score of 64.17% can correctly identify L1 speakers 88.75% of the time and L2 speakers 81.25% of the time. AUC = area under the curve; ROC = receiver operating characteristic.

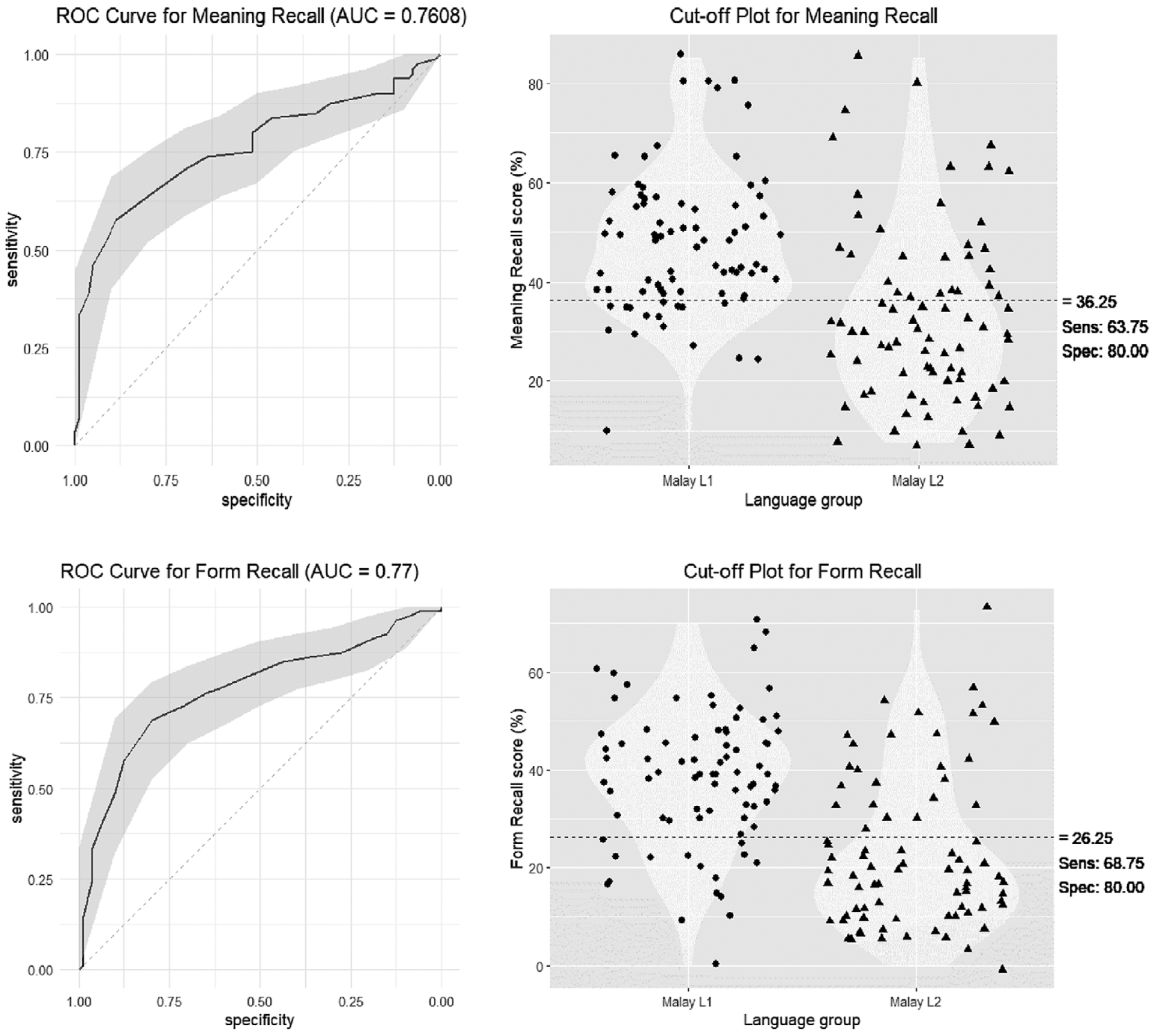

Figure 4. Continued.

significant proportion of variance in the yes/no vocabulary test scores. As recall of word meanings is required for many receptive tasks such as listening and reading (Nation, 2020; Schmitt, 2010), it is not surprising to observe unique predictions from Meaning Recall test scores given that yes/no lexical decision task is fundamentally a receptive task. On the other hand, Form Recall test scores did not explain additional unique variance in LexMAL scores because recall of word forms is usually required only for productive tasks such as speaking and writing (Nation, 2020; Schmitt, 2010). Taken together, these findings suggest that yes/no vocabulary test scores could be a reliable indicator of bilinguals' receptive lexical proficiency (in line with Mochida & Harrington, 2006) because test takers' performance in the test corresponded well with their knowledge of meaning recognition, form recognition and meaning recall.

At the item level, the generalized mixed-effects model revealed that different aspects of form-meaning knowledge were found to influence item accuracy in the vocabulary tests. Items that were correctly identified in LexMAL were more likely to be answered correctly in the Form Recognition test (as indicated by positive log odds), but less likely to be answered correctly in the Meaning Recall and Form Recall tests (as indicated by negative log odds; see Figure 3). The higher tendency for participants to recognize the word forms in the Form Recognition test following their correct identification in LexMAL suggests that form recognition knowledge supported their ability to identify them as real words in the yes/no vocabulary test. However, for these LexMAL items that were identified as real words, participants may not be able to recall their meanings or the word forms when their meanings were provided. Furthermore, correct identification of words in LexMAL does not indicate that test takers would be able to recognize their meanings given the word forms. Therefore, researchers who use yes/no vocabulary tests should be made aware of this limitation of the vocabulary knowledge measured and be cautious not to overclaim participants' mastery of the vocabulary items. Nevertheless, our findings still support the use of yes/no vocabulary tests as a lexical proficiency test because its item accuracy corresponds well with participants' form recognition knowledge (Elgort, 2013; McLean et al., 2020).

The generalized mixed-effects model also revealed a significant difference in form-meaning knowledge between the two language groups because Malay L1 speakers outperformed L2 speakers across all vocabulary tests. This is consistent with previous studies that reported L1 speakers to have larger vocabulary sizes than L2 speakers (Rahman et al., 2018). The L1–L2 speaker difference has also been consistently demonstrated in previous yes/no vocabulary test validation studies (Amenta et al., 2020; Brysbaert, 2013; Izura

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

et al., 2014; Lee et al., 2023; Salmela et al., 2021; Wen et al., 2023), providing support for the validity of yes/no vocabulary tests as a lexical proficiency measure that can discriminate between L1 and L2 speakers.

It may seem surprising that even the highly proficient L1 speakers obtained low scores in Meaning Recall and Form Recall tests (see Table 3). This is however in-line with the findings of our pilot study. The reason for the low scores is two-fold. As observed also in previous studies, recall tasks are more difficult than recognition tasks and bilinguals usually score lower in the former because recall tasks do not provide choices, and most importantly, they do not account for partial knowledge (González-Fernández & Schmitt, 2020; Laufer & Aviad-Levitzky, 2017; Laufer & Goldstein, 2004; McLean et al., 2020; Stewart et al., 2023). Furthermore, the LexMAL target words were carefully selected to be difficult enough even for the L1 speakers to capture the variation in vocabulary knowledge of highly proficient L1 speakers (M_{Zipf} = 3.56, SD_{Zipf} = 0.54; see Lee et al., 2023). In addition to having a good blend of lexical decision difficulty (M_{accuracy} = 48.41%, SD_{accuracy} = 26.30%; taken from Yap et al., 2010), 50 out of the 60 target words (83.33%) have <50% translation accuracy by L1 speakers (M_{accuracy} = 24.44%, SD_{accuracy} = 21.95%; taken from Lee et al., 2022).

As the same target words from LexMAL were tested across the four form-meaning levels, the low accuracy in the Form Recall and Meaning Recall tests of the L1 speakers can be attributed to the difficulty level of the tasks. Recognizing the form and/or meaning of these words was easier for the L1 speakers when they were prompted by cues (e.g., recognizing the answer among foils), suggesting that they know these vocabulary items to some extent (i.e., partial knowledge; Laufer & Aviad-Levitzky, 2017). In contrast, recalling the form and/or meaning of the vocabulary items was more difficult when they appeared in isolation or a clueless context, even for the highly proficient L1 speakers. This finding suggests that mastery of recognition knowledge precedes that of recall (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004) and individual differences in these distinct aspects of form-meaning knowledge can still be heterogenous among the highly proficient L1 speakers (see Figure 4), further indicating the importance of measuring L1 lexical proficiency in research (Brysbaert et al., 2016; Hulstijn, 2015; Lee et al., 2022). Vocabulary tests like LexMAL (Lee et al., 2023), for example, could serve as a good tool for language research to measure the lexical proficiency of L1 and L2 speakers of the target language on the same scale.

In terms of test discrimination ability, the ROC curve analyses revealed that LexMAL and the recognition tests had the highest discriminant ability (i.e., sensitivity and specificity of at least 80%) in identifying L1 and L2 speakers. This could be because LexMAL and the recognition tests were easier for L1 speakers than L2 speakers; therefore, the L1 speakers consistently scored higher than the cutoff scores compared to L2 speakers. The Meaning Recall and Form Recall tests, on the other hand, showed weaker discrimination between L1 and L2 speakers (AUC < 0.80) and identification of L1 speakers based on vocabulary knowledge (sensitivity < 70%). In addition to the considerably higher difficulty of the recall tests than recognition tests (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004), the difficulty level of the vocabulary items (as indicated by Lee et al., 2022; see discussion above) even for the L1 speakers also contributes to the great variation of performance among the L1 speakers and a good number of L1 speakers scoring below the optimal cutoff scores. Taken together, our findings suggest that yes/no vocabulary and recognition tests are better options than recall tests when the purpose is to distinguish the form-meaning vocabulary knowledge of L1 and L2 speakers or to identify speakers from a specific speaker group.

Whereas all vocabulary tests in the present study displayed high reliability and good discrimination ability, the suitability of the research tests depends on the purpose of the testing. For instance, if the purpose is to measure lexical proficiency, using one of these vocabulary tests might be sufficient because their scores were highly correlated. Given their robust correlation, form-meaning recognition tests (Meaning Recognition or Form Recognition tests) could be a better option than yes/no vocabulary tests when a direct demonstration of vocabulary knowledge is required (e.g., to demonstrate word knowledge at a specific level) while accounting for partial knowledge (Laufer & Aviad-Levitzky, 2017). It is noteworthy, however, that the L1 speakers might demonstrate ceiling performance in the form-meaning recognition tests compared to the yes/no vocabulary tests. This could potentially be due to the cues presented in the test stimuli. In cases where direct demonstration of vocabulary knowledge is needed without accounting for partial knowledge, the form-meaning recall tests (Meaning Recall or Form Recall tests) can be useful (McLean et al., 2020).

If the language testing purpose is to distinguish between L1 and L2 speakers, and at the same time capture a good variation in both groups of speakers, recognition tests appeared to be a better option than recall tests. Recognition tests are easier than recall tests and therefore require less task demands on participants. Specifically, the yes/no vocabulary test offers a quick and valid measure of lexical proficiency. In contrast to the form-meaning recognition tests, the yes/no vocabulary tests are easier to construct and more items can be tested within a short period of time. As test scores from the yes/no vocabulary test were positively predicted by form recognition but not meaning recognition knowledge, the test scores from the yes/no vocabulary test to some extent capture test takers' ability to recognize some real word forms, even though they may not recognize the word meanings.

The present study provides evidence that the yes/no vocabulary test is effective in distinguishing L1 and L2 speakers, and captures form recognition knowledge to some extent, which is useful for test score interpretation in research that seeks a quick proficiency test. Our findings, however, do not suggest that it is superior to other form-meaning vocabulary tests, nor do they imply that it could serve as a replacement for these tests. For detailed assessments and research that seeks measurements at specific form-meaning knowledge levels, form-meaning vocabulary tests are useful if these tests are available in the target language. Future research could use factor analyses to explore the structure underlying yes/no vocabulary tests and form-meaning vocabulary tests to gain an understanding of the constructs measured by different vocabulary tests.

The predictive ability of yes/no and other form-meaning vocabulary tests for reading comprehension was not assessed in the present study. Therefore, our findings are unable to provide information about the best vocabulary test to predict reading performance (see Laufer & Aviad-Levitzky, 2017 and McLean et al., 2020, for this line of investigation). Future research is needed to investigate how different vocabulary tests could predict different linguistic tasks, such as word recognition and reading comprehension. As far as we are aware, there is no standardized reading comprehension test available for Malay; therefore, future research could consider creating or adapting a reading comprehension test from existing literature (e.g., Siegelman et al., 2022) to further explore the relationship between various vocabulary tests and reading comprehension performance in the Malay language.

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press

5. Conclusions

The present study used four form-meaning vocabulary tests to evaluate the contribution of bilinguals' form-meaning knowledge to their language proficiency as measured by a yes/no vocabulary test. Bilinguals' form-meaning knowledge explained a significant proportion of the variance in their yes/no vocabulary test scores, with knowledge of form recognition being the best predictor, followed by meaning recognition and meaning recall. Furthermore, our results suggest that yes/no vocabulary tests primarily assess recognition knowledge, and those who correctly identify the test items are also more likely to recognize the word forms given their meanings. However, participants may not be able to recall these test items' meanings or word forms given their meanings. Importantly, LexMAL and recognition tests were found to be more effective than recall tests in distinguishing between L1 and L2 speakers' form-meaning vocabulary knowledge. With meaning recognition, form recognition and meaning recall serving as predictors of LexMAL scores, and form recognition being the positive predictor of item accuracy in LexMAL, our study provides evidence to support the use of yes/no vocabulary tests as quick and reliable lexical proficiency measures to estimate bilinguals' receptive language proficiency.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/S1366728924001007.

Data availability statement. The data collected during this study are available on the Open Science Framework repository at https://osf.io/9nyqd/.

Competing interests. The author(s) declare none.

Acknowledgements. The authors thank Khadijah Khalid and Norashiqin Shaari for helpful discussions about the stimuli writing, as well as Siti Amirah and Zhi Diong for the inter-rater scoring.

References

Amenta, S., **Badan, L.**, & **Brysbaert, M.** (2021). LexITA: A quick and reliable assessment tool for Italian L2 receptive vocabulary size. *Applied Linguistics*, **42**(2), 292–314. https://doi.org/10.1093/applin/amaa020

Anderson, R. C., & **Freebody, P.** (1983). Reading comprehension and the assessment and acquisition of word knowledge. *Advances in Reading Language Research*, **2**, 231–256.

Aviad-Levitzky, T., **Laufer, B.**, & **Goldstein, Z.** (2019). The New Computer Adaptive Test of Size and Strength (CATSS): Development and validation. *Language Assessment Quarterly*, **16**(3), 345–368. https://doi.org/10.1080/15434303.2019.1649409

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

Brysbaert, M. (2013). Lextale_FR: A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica*, **53**(1), 23–37. https://doi.org/10.5334/pb-53-1-23

Brysbaert, M., **Lagrou, E.**, & **Stevens, M.** (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. Bilingualism *(Cambridge, England)*, **20**(3), 530–548. https://doi.org/10.1017/S1366728916000353

Brysbaert, M., **Stevens, M.**, **Mandera, P.**, & **Keuleers, E.** (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, **7**, 1116–1116. https://doi.org/10.3389/fpsyg.2016.01116

Brysbaert, M., **Mandera, P.**, & **Keuleers, E.** (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, **27**(1), 45–50. https://doi.org/10.1177/0963721417727521

Chan, I. L., & **Chang, C. B.** (2018). LEXTALE_CH: A quick, characterbased proficiency test for Mandarin Chinese. Proceedings of the Annual Boston University Conference on Language Development, **42**(1), 114–130. https://hdl.handle.net/2144/29734

Diedenhofen, B. & **Musch, J.** (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, **10**(4): e0121945. https://doi.org/10.1371/journal.pone.0121945

Diependaele, K., **Lemhöfer, K.**, & **Brysbaert, M.** (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*. **66**, 843–863. https://doi.org/10.1080/17470218.2012.720994

Elgort. I. (2013). Effects of LI definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, **30**(2), 253–272. https://doi.org/10.1177/0265532212459028

Fromont, L., **Royle, P.**, & **Steinhauer, K.** (2020). Growing Random Forests reveals that exposure and proficiency best account for individual variability in L2 (and L1) brain potentials for syntax and semantics. *Brain and Language*, **204**, 104770–104770. https://doi.org/10.1016/j.bandl.2020.104770

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

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

Harrington, M. (2018). L2 word recognition skill and its measurement. In *Lexical facility: Size, recognition speed and consistency as dimensions of second language vocabulary knowledge* (pp. 45–65). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-37262-8_3

Holmes, V. M. (2009). Bottom-up processing and reading comprehension in experienced adult readers. *Journal of Research in Reading*, **32**(3), 309–326. https://doi.org/10.1111/j.1467-9817.2009.01396.x

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, **15**(2), 422–433. https://doi.org/10.1017/S1366728911000678

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.

Ibrahim, S. (Ed.). (2002). Kamus Dwibahasa: Bahasa Inggeris-Bahasa Melayu *(Edisi Kedua)*. Dewan Bahasa dan Pustaka.

Izura, C., **Cuetos, F.**, & **Brysbaert, M.** (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica (Valencia)*, **35**(1), 49–66.

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

Lalkhen, A. G., & **McCluskey, A.** (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain*, **8**(6), 221–223. https://doi.org/10.1093/bjaceaccp/mkn041

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**(4), 729–741. https://doi.org/10.1111/modl.12431

Laufer, B., & **Goldstein, Z.** (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, **54**(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Lee, L. C., **Rickard Liow, S. J.**, & **Wee, M.-L. O.** (2007). Morphological structure of Malay: Using psycholinguistic analyses of rated familiarity. In M. Alves, P. Sidwell, & D. Gil (Eds.), *SEALSVIII: Papers from the 8th meeting of the Southeast Asian Linguistics Society* (pp. 109–119). Pacific Linguistics.

Lee, S. T., **van Heuven, W. J.**, **Price, J. M.**, & **Leong, C. X. R.** (2022). Translation norms for Malay and English words: The effects of word class, semantic variability, lexical characteristics, and language proficiency on translation. *Behavior Research Methods*, **55**, 3585–3601. https://doi.org/10.3758/s13428-022-01977-3

Lee, S. T., **van Heuven, W. J. B.**, **Price, J. M.**, & **Leong, C. X. R.** (2023). LexMAL: A quick and reliable lexical test for Malay speakers. *Behavior Research Methods*, **56**, 4563–4581. https://doi.org/10.3758/s13428-023-02202-5

Lemhöfer, K., & **Broersma, M.** (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, **44**(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka least-squares means.* R package version 1.8.5. https://CRAN.R-project.org/package=emmeans

Li, P., **Zhang, F.**, **Yu, A.**, & **Zhao, X.** (2019). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. In *Bilingualism* (pp. 1–7). Cambridge University Press. https://doi.org/10.1017/S1366728918001153

Mainz, N., **Shao, Z.**, **Brysbaert, M.**, & **Meyer, A. S.** (2017). Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. *Frontiers in Psychology*, **8**, 1164–1164. https://doi.org/10.3389/fpsyg.2017.01164

Masrai, A. (2022). The development and validation of a Lemma-based yes/no vocabulary size test. *SAGE Open*, **12**(1), 215824402210743. https://doi.org/10.1177/21582440221074355

Mazlan, I. R., **Hassnan, N. M.**, & **Rusli, Y. A.** (2024). A comparison of narrative abilities in Malay school-age typically developing children and children with developmental language disorder. *Clinical Linguistics & Phonetics*, 1–22. https://doi.org/10.1080/02699206.2024.2359462

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

Meara, P., & **Miralpeix, I.** (2016). *Tools for Researching Vocabulary*. Multilingual Matters. https://doi.org/10.21832/9781783096473

Meara, P. M., & **Jones, G.** (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society* (pp. 80–87). CILT.

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

Mohamed, M. M., & **Jared, D.** (2024). Malay Lexicon Project 3: The impact of orthographic–semantic consistency on lexical decision latencies. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218241234668

Mohamed, M. M., **Yap, M. J.**, **Chee, Q. W.**, & **Jared, D.** (2023). Malay Lexicon Project 2: Morphology in Malay word recognition. *Memory & Cognition*, **51** (3), 647–665. https://doi.org/10.3758/s13421-022-01337-8

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, **5**(1), 12–25.

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

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

Nation, I. S. P. (2020). The different aspects of vocabulary knowledge. In *The Routledge handbook of vocabulary studies* (1st ed., pp. 15–29). Routledge. https://doi.org/10.4324/9780429291586-2

Nation, I. S. P. (2022). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P., & **Beglar, D.** (2007) A vocabulary size test. *The Language Teacher*, **31**(7), 9–13.

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

Nomoto, H., **Choi, H.**, **Moeljadi, D.**, & **Bond, F.** (2018). MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. In *Proceedings of the LREC 2018 Workshop "The 13th Workshop on Asian Language Resources"*. European Language Resources Association (ELRA): Miyazaki, Japan (pp. 36–43).

Park, H., **Solon, M.**, **Dehghan-Chaleshtori, M.**, & **Ghanbar, H.** (2022). Proficiency reporting practices in research on second language acquisition: Have we made any progress? *Language Learning*, **72**(1), 198–236. https://doi.org/10.1111/lang.12475

Puig-Mayenco, E., **Chaouch-Orozco, A.**, **Liu, H.**, & **Martín-Villena, F.** (2023). The LexTALE as a measure of L2 global proficiency: A cautionary tale based on a partial replication of Lemhöfer and Broersma (2012). *Linguistic Approaches to Bilingualism*, **13**(3), 299–314. https://doi.org/10.1075/lab.22048.pui

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

Rahman, A., **Yap, N. T.**, & **Darmi, R.** (2018). The association between vocabulary size and language dominance of bilingual Malay-English undergraduate, *3 L: Language, Linguistics, Literature the South East Asian Journal of English Language Studies*, **24**(4), 85–101. https://doi.org/10.17576/3L-2018-2404-07

Raven, J. (2000). The Raven's progressive matrices: Change and Stability over Culture and time. *Cognitive Psychology*, **41**(1), 1–48. https://doi.org/10.1006/cogp.1999.0735

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

Read, J. P., **Haas, A. L.**, **Radomski, S.**, **Wickham, R. E.**, & **Borish, S. E.** (2015). Identification of hazardous drinking with the young adult alcohol consequences questionnaire: Relative operating characteristics as a function of gender. *Psychological Assessment*, **28**(10), 1276–1289. https://doi.org/10.1037/pas0000251

Robin, X., **Turck, N.**, **Hainard, A.**, **Tiberti, N.**, **Lisacek, F.**, **Sanchez, J. C.**, & **Muller, M.** (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**(1), 77–77. https://doi.org/10.1186/1471-2105-12-77

R Core Team (2021). R: A language and environment for statistical computing. Published online 2020. *Supplemental Information References S*, **1**, 371–78.

Rusli, A. G., **Mohamed Husin, N.**, & **Chin, L. Y.** (2006). Pangkalan data korpus DBP: Perancangan, pembinaan dan pemanfaatan. In Z. Ahmad (Ed.), *Aspek nahu praktis Bahasa Melayu* (pp. 21–25). Bangi: Universiti Kebangsaan Malaysia Press.

Salmela, R., **Lehtonen, M.**, **Garusi, S.**, & **Bertram, R.** (2021). Lexize: A test to quickly assess vocabulary knowledge in Finnish. *Scandinavian Journal of Psychology*, **62**(6), 806–819. https://doi.org/10.1111/sjop.12768

Sarrett, M., **Shea, C.**, & **McMurray, B.** (2022). Within- and between-language competition in adult second language learners: Implications for language proficiency. *Language, Cognition and Neuroscience*, **37**(2), 165–181. https://doi.org/10.1080/23273798.2021.1952283

Schmitt, N. (2010). Issues of vocabulary acquisition and use. In: *Researching vocabulary. Research and practice in applied linguistics*. Palgrave Macmillan. https://doi.org/10.1057/9780230293977_2

Schmitt, N., **Cobb, T.**, **Horst, M.**, & **Schmitt, D.** (2015). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). *Language Teaching*, **50**(2), 212–226. https://doi.org/10.1017/S0261444815000075

Schmitt, N., **Nation, P.**, & **Kremmel, B.** (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, **53**(1), 109–120. https://doi.org/10.1017/S0261444819000326

Schmitt, N., **Schmitt, D.**, & **Clapham, C.** (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, **18**(1), 55–88.

Siegelman, N., **Elgort, I.**, **Brysbaert, M.**, **Agrawal, N.**, **Amenta, S.**, **Arsenijević Mijalković, J.**, **Chang, C. S.**, **Chernova, D.**, **Chetail, F.**, **Clarke, A. J. B.**, **Content, A.**, **Crepaldi, D.**, **Davaabold, N.**, **Delgersuren, S.**, **Deutsch, A.**, **Dibrova, V.**, **Drieghe, D.**, **Đurđević, D. F.**, **Finch, B.**, **Frost, R.**, **Gattei, C. A.**, **Geva, E.**, **Godfroid, A.**, **Griener, L.**, **Hernández-Rivera, E.**, **Ivanenko, A.**, **Järvikivi, J.**, **Kawaletz, L.**, **Khare, A.**, **Lee, J. R.**, **Lee, C. E.**, **Manouilidou, C.**, **Marelli, M.**, **Mashanlo, T.**, **Mišić, K.**, **Miwa, K.**, **Palma, P.**, **Plag, I.**, **Rezanova, Z.**, **Riimed, E.**, **Rueckl, J.**, **Schroeder, S.**, **Sekerina, I. A.**, **Shalom, D. E.**, **Slioussar, N.**, **Slosar, N. M.**, **Taler, V.**, **Thériault, K.**, **Titone, D.**, **Tumee, O.**, **van de Wetering, R.**, **Verma, A.**, **Weiss, A. F.**, **Wu, D. H.**, & **Kuperman, V.** (2024). Rethinking first language–second language similarities and differences in English proficiency: Insights from the ENglish Reading Online (ENRO) Project. *Language Learning*, **74**(1), 249– 294. https://doi.org/10.1111/lang.12586

Siegelman, N., **Schroeder, S.**, **Acartürk, C.**, **Ahn, H.-D.**, **Alexeeva, S.**, **Amenta, S.**, **Bertram, R.**, **Bonandrini, R.**, **Brysbaert, M.**, **Chernova, D.**, **Da Fonseca, S. M.**, **Dirix, N.**, **Duyck, W.**, **Fella, A.**, **Frost, R.**, **Gattei, C. A.**, **Kalaitzi, A.**, **Kwon, N.**, **Lõo, K.**, **Marelli, M.**, **Papadopoulos, T. C.**, **Protopapas, A.**, **Savo, S.**, **Shalom, D. E.**, **Slioussar, N.**, **Stein, R.**, **Sui, L.**, **Taboh, A.**, **Tønnesen, V.**, **Usal, K. A.**, & **Kuperman, V.** (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, **54**(6), 2843–2863. https://doi.org/10.3758/s13428-021-01772-6

Singh, A., **Wang, M.**, & **Faroqi-Shah, Y.** (2022). The influence of romanizing a non-alphabetic L1 on L2 reading: the case of Hindi–English visual word recognition. *Reading & Writing*, **35**(6), 1475–1496. https://doi.org/10.1007/s11145-021-10241-7

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, **87**(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Stewart, J., **Gyllstad, H.**, **Nicklin, C.**, & **McLean, S.** (2023). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*, **41**(1), 89–108. https://doi.org/10.1177/02655322231162853

Surrain, S., & **Luk, G.** (2019). Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism: Language and Cognition*, **22**(2), 401–415. https://doi.org/10.1017/S1366728917000682

Tosun, S., **Filipović, L.** (2022). Lost in translation, apparently: Bilingual language processing of evidentiality in a Turkish–English Translation and judgment task. *Bilingualism: Language and Cognition*, **25**, 739–754. https://doi.org/10.1017/S1366728922000141

Treffers-Daller, J. (2019). What defines language dominance in bilinguals? *Annual Review of Linguistics*, **5**(1), 375–393. https://doi.org/10.1146/annurev-linguistics-011817-045554

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, **33**, 339–372. https://doi.org/10.1017/S0272263111000015

van Heuven, W. J. B., **Mandera, P.**, **Keuleers, E.**, & **Brysbaert, M.** (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, **67**(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Webb, S., **Sasao, Y.**, & **Ballance, O.** (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

Wen, Y., **Qiu, Y.**, **Leong, C. X. R.**, & **van Heuven, W. J. B.** (2023). LexCHI: A quick lexical test for estimating language proficiency in Chinese. *Behavior Research Methods*, **56**, 2333–2352. https://doi.org/10.3758/s13428-023-02151-z

Wen, Y., & **van Heuven, W. J. B.** (2017). Non-cognate translation priming in masked priming lexical decision experiments: A meta-analysis. *Psychonomic Bulletin & Review*, **24**(3), 879–886. https://doi.org/10.3758/s13423-016-1151-1

Yap, M. J., **Liow, S. J. R.**, **Jalil, S. B.**, & **Faizal, S. S. B.** (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, **42**(4), 992–1003. https://doi.org/10.3758/BRM.42.4.992

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

Zhou, C., & **Li, X.** (2022). LextPT: A reliable and efficient vocabulary size test for L2 Portuguese proficiency. *Behavior Research Methods*, **54**(6), 2625–2639. https://doi.org/10.3758/s13428-021-01731-1

https://doi.org/10.1017/S1366728924001007 Published online by Cambridge University Press