University of Nottingham
UK | CHINA | MALAYSIA

MEASURING LANGUAGE PROFICIENCY IN MALAY FIRST AND SECOND LANGUAGE SPEAKERS

Soon Tat Lee

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

in Psychology

April 2024

Acknowledgement

I would like to express my heartfelt gratitude to my supervisors, Dr. Christine Leong, Dr. Jessica Price, and Dr. Walter van Heuven, for their invaluable guidance and unwavering support throughout my PhD journey. Your mentorship has not only shaped my research but also my growth as a scholar. The wisdom you have imparted, the scholarly insights you have shared, and the commitment you have shown have been nothing short of transformative. This endeavour would not be possible without your dedication and support.

I extend my sincere appreciation to Dr. Csaba Szabo for his invaluable feedback on my PhD projects. Your expertise and insights have played a pivotal role in enhancing the depth of my research. Special thanks to Dr. Low Hui Min for her help in participant recruitment and to all the research participants for their time in participating in my studies.

To my dear friends, Bryan, Jasmine, Kai Hao, Keith, Kelly, Mei Ling, Josh, Tijn, and Tsuey Bing, thank you for being the remarkable friends that you are. You have been the pillars of unwavering support, the beacons of laughter during the darkest nights, and the bearers of endless encouragement. Your friendship has made this journey all the more fulfilling.

From the depths of my heart, I wish to thank my family for their unconditional love and support throughout this journey. Your belief in me has been my pillar of strength, and I owe a significant part of my success to you.

Table of Content

List of Acronyms

AUC	Area under the curve
BT	Backward translation
FT	Forward translation
IELTS	International English Language Testing System
IRT	Item response theory
TOEFL	Test of English as a Foreign Language
L1	First language
L2	Second language
LexMAL	Lexical test for Malay Speakers
LexTALE	Lexical Test for Advanced Learners of English
ROC	Receiver operator characteristic

List of Tables

List of Figures

Abstract

Measuring language proficiency is essential in linguistics and psycholinguistics research that investigate bilingual language processing (e.g., Fromont et al., 2020; Sarrett et al., 2022; Singh et al., 2022; Tosun & Filipović, 2022) and cognitive control (e.g., Luque & Morgan-Short, 2021; Ning, 2021). Despite often being construed as a moderating variable in bilingual research, there is a great variability in how language proficiency is operationalised and measured (Hulstijn, 2015; Puig-Mayenco et al., 2023; Surrain & Luk, 2019; Treffers-Daller, 2019; Tremblay, 2011). A systematic review on second language acquisition research (Park et al., 2022) revealed that about 58% of the studies estimated second language proficiency using variables that were closely related to language proficiency (e.g., years of instruction, self-rated proficiency). In contrast, only 42% of them assessed language proficiency with a test (e.g., validated vocabulary tests). Despite being shown as a more consistent measure, and importantly, correlated well with other language performances (e.g., Diependaele et al., 2013; Lemhöfer & Broersma, 2012; Wen & van Heuven, 2017b), implementation of language proficiency test is not always feasible because existing language proficiency tests might not be available in understudied languages or could be impractical (e.g., too time-consuming) for experimental settings. For instance, there is currently no freely available quick Malay proficiency test, although there are 377 million Malay speakers in the world.

This thesis aimed to improve the methodological rigour of bilingual language testing through the development of psycholinguistic tools for Malay, an understudied language. As a starting point, this thesis aimed to create a large database of Malay and English bidirectional translation norms to facilitate Malay-English cross-linguistic stimulus selection. Malay-English bidirectional translation equivalents were collected from highly proficient bilinguals. The second aim of this thesis was to develop and validate the Lexical Test for Malay Speakers (LexMAL), a vocabulary test that can distinguish Malay learners of various proficiency levels while assessing their discrete vocabulary knowledge based on a selected list of target words presented in isolation (i.e., out of context). External criterion measures of language proficiency were used to validate the test, including translation tasks, the cloze test, and self-rated proficiency. Subsequently, this thesis sought to justify the utility of the widely used yes/no vocabulary test scores to gauge bilinguals' lexical proficiency. Using LexMAL and four newly created form-meaning vocabulary tests, the relationships between form-meaning vocabulary knowledge and yes/no vocabulary test scores were examined in depth.

Chapter 2 of this thesis describes the development of the Malay-English and English-Malay translation norms through forward and backward translation tasks. Information from these translation norms guided the selection of appropriate translation stimuli used in Chapter 3 to assess language proficiency. The Malay and English translation norms presented are among the first collected from highly proficient Malay-English bilinguals. The study also investigated Malay-English translation ambiguity in bidirectional translation tasks and how it was affected by other lexical characteristics (e.g., word class or frequency of occurrence). The study gathered English translations of 1,004 Malay words and Malay translations of 845 English words. The study revealed high prevalence of translation ambiguity between the Malay and English languages. The findings corroborate that translation ambiguity could emerge due to the conceptual and morphological mapping differences between the target and source languages, as well as language specific properties of the language pairs under investigation (Schwieter & Prior, 2020). Together with lexical and semantic information of the source and target words, these norms could be useful references to aid stimuli selection for future experimental studies (e.g., Jouravlev & Jared, 2020) and computer simulations (e.g., Dijkstra et al., 2019). Serving as the first freely available translation norms database for researchers conducting language research with Malay-English bilinguals, this study is published in *Behavior Research Methods* (i.e., S. T. Lee et al., 2022).

Chapter 3 of this thesis presents the construction process of LexMAL, which builds on the Malay-English translation norms. The development of LexMAL began with the stimuli writing and selection, pilot testing of the LexMAL prototype, item assessment and ended with validation of the final LexMAL. To estimate language proficiency for Malay first language (L1) and second language (L2) speakers, an initial 180-items LexMAL prototype was evaluated on 60 Malay L1 and 60 L2 speakers. Sixty words and thirty nonwords with the highest discriminative power that span across the full difficulty range were selected for the final LexMAL. The validity of LexMAL was established by demonstrating reliable discrimination between L1 and L2 speakers, and significant correlations with other Malay language tasks. Importantly, LexMAL outperformed self-rated proficiency in the correlations with cloze task performance, indicating that objective measures like LexMAL are better estimates of language proficiency than subjective self-ratings (Khare et al., 2013; Lemhöfer & Broersma, 2012; Tomoschuk et al., 2019; Wen & van Heuven, 2017a). As far as we are aware, LexMAL is the first validated Malay lexical test that can reliably measure the proficiency of L1 and L2 speakers. LexMAL is useful for researchers in, for example, linguistics, psychology, and education that require a quick (less than 5 minutes), practical and objective proficiency measure. LexMAL can be taken online at https://www.lexmal.org/, or a paper and pencil version of LexMAL can be downloaded from https://osf.io/8y4ft/. The paper is published in *Behavior Research Methods* (i.e., S. T. Lee et al., 2023).

Chapter 4 of this thesis investigates the prediction of bilinguals' form-meaning vocabulary knowledge to their item accuracy on LexMAL. Four vocabulary tests were developed to assess bilinguals' knowledge of meaning recognition, form recognition, meaning recall, and form recall. The study found that language dominance affected the form-meaning knowledge of bilinguals, with L1 speakers performing better than L2 speakers. Additionally, the accuracy of Meaning Recognition, Form Recognition, and Meaning Recall tests explained 59% of the variance in LexMAL scores. 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 score, and form recognition being the positive predictor of item accuracy in LexMAL, this 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. The paper has been submitted to Bilingualism: Language and Cognition journal and is currently under review.

Chapter 5 summarises and discusses the main findings of the thesis. The theoretical contributions of these findings are discussed in relation to the current understanding of test development and validation, bilingual language processing, and vocabulary testing. Additionally, this discussion critically evaluates the limitations of the studies while proposing potential directions for future research. Overall, the three empirical studies in this thesis advance current knowledge in bilingual language testing by assessing the vocabulary knowledge of the L1 and L2 speakers on the same scale. The psycholinguistic tools presented in this thesis enable researchers to make informed decisions when selecting lexical stimuli and lexical test formats, as well as interpret research findings based on accurate measures of participants' language proficiency.

Chapter 1

Introduction

This chapter reviews language proficiency models and discusses how language proficiency testing can be informed by the existing theories to develop a valid test that could reliably measure the language proficiency of Malay speakers. Under the overarching nature of language proficiency construct, vocabulary knowledge is presented as one of the fundamental constructs that can be measured to estimate language proficiency. This chapter dives into the rich variety of commonly used vocabulary tests, and illustrates how vocabulary tests can serve as a reliable and practical language proficiency measure in research. Finally, this chapter provides an overview of the diverse multilingual context in Malaysia and underscores the distinctive role of the Malay language within bilingual research. The practical implication of developing a language proficiency test tailored to Malay-speaking bilinguals is also discussed.

1.1 Language Proficiency Models

Language proficiency is considered to be an important variable for theories of bilingual language processing. There is, however, a great variability in how it is operationalised and measured (Park et al., 2022; Puig-Mayenco et al., 2023; Surrain & Luk, 2019; Treffers-Daller, 2019; Tremblay, 2011). For instance, language proficiency can be defined as narrowly as the extent to which a language is known in terms of syntax, morphology, phonology, phonetics, and vocabulary (e.g., grammatical knowledge; Treffers-Daller, 2019), or as broadly as a language ability construct that includes language knowledge (a combination of grammatical, textual, functional and sociolinguistic knowledge) and strategic knowledge (metacognitive strategies, such as goal setting and planning) (Bachman & Palmer, 1996; 2010). Given the highly complex and multidimensional nature of language proficiency, it is important to operationalise the language proficiency model and components that underlie the construct before deciding how it can be assessed (Bialystok, 2001; Hulstijn, 2015; Park et al., 2022).

When it comes to assessing bilinguals’ language proficiency, language is frequently discussed as if it can be objectively measured, and the acquisition of language is viewed as a predictable progression from not knowing to complete mastery (Bialystok, 2001). However, the standards used to assess language proficiency are rarely specified, even though they are fundamental to language research and individual language ability. What, for instance, qualifies as typical language proficiency? How do we define language proficiency, its elements, and the range of variation that is acceptable? To organise this complex and multidimensional construct of language proficiency into a coherent statement about bilinguals' ability to learn and use language, the Analysis and Control Framework (Bialystok, 2001) and the Theory of Basic and Higher Language Cognition (Hulstijn, 2015) offer some identifiable cognitive operations to operationalise bilinguals’ language proficiency. These frameworks attempt to define the boundaries of language proficiency, acknowledge variability among bilinguals and specify various dimensions of determining language proficiency. The following subsections briefly summarise the two frameworks.

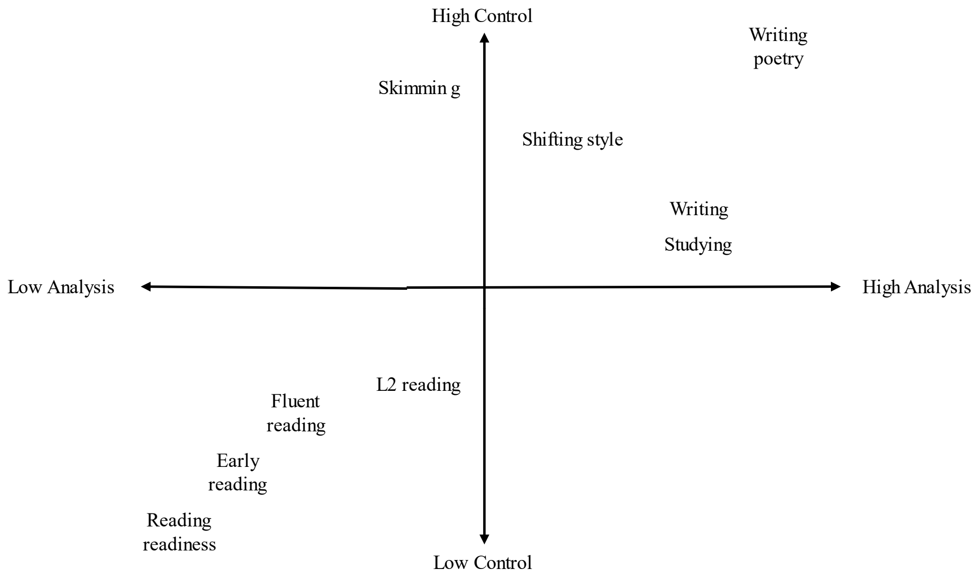
1.1.1 Analysis and Control Framework (Bialystok, 2001)

The language proficiency model by Bialystok (2001) describes the intricate relationship between bilinguals' linguistic competency and cognitive control. It has been used to investigate the linguistic and metalinguistic development in children growing up with one or several languages (e.g., Spit et al., 2023), and in adolescents and adults acquiring a second language (e.g., Dash & Kar, 2020). According to Bialystok's (2001) framework, language proficiency can be defined as the ability to perform a language task that requires specific linguistic and cognitive demands to a specific level of performance indicated by either an objective criterion or a normative standard. Within the framework, the linguistic and cognitive demands of a language task are identified by comparing the relative reliance on the analysis of representational structure and control of attention (see Figure 1.1 for the increasing demands on analysis and control from reading to writing). Analysis of representational structure refers to the process by which bilinguals access and use their linguistic knowledge, whereas control of attention refers to the cognitive processes that bilinguals use to manage the competition between their language systems and control the cognitive processes related to language use. Depending on the levels of cognitive demands on the two processes, the level of language proficiency required to perform in the specific language task differs accordingly. For example, writing poetry might require a high level of linguistic knowledge and cognitive control, while fluent reading might place relatively lower demands on linguistic knowledge and cognitive control.

Even when the language tasks fall under the same language domain (such as literacy; see Figure 1.1 for the demands required for each literacy task), their linguistic and cognitive demands can differ strikingly. For instance, literacy tasks can span across diverse modalities (e.g., early or fluent reading), purposes (e.g., studying or skimming), and genres (e.g., fiction or poetry). Each of these tasks would involve Analysis and Control to different extents, and therefore convey different levels of language proficiency. A language learner who performs well in a skimming task might not perform as well in a study task because the former requires a higher level of attention to perform, whereas the latter requires a higher level of formal knowledge. Therefore, to construct appropriate language proficiency measures for a specific research question (e.g., academic reading comprehension), it becomes important to have these underlying cognitive demands clearly specified.

Figure 1. 1

Domains of language use in Analysis and Control Framework (Bialystok, 2001)



Note. Adapted from Bialystok (2001, pp. 17)

According to the Analysis and Control Framework, different language tasks would involve different linguistic and cognitive demands. It proposes a need for a rigorous test development approach by highlighting the importance of accurately identifying the linguistic and cognitive demands in language tasks used for language proficiency testing. Because linguistic and cognitive processing are prerequisites for language proficiency, it is first necessary to determine the components of cognitive processes that can be used to measure language proficiency (e.g., storage of vocabulary knowledge). In this thesis, this was done by establishing the criterion-referenced performances of Malay language learners in various language proficiency tests. This important first step serves as the foundation for future development of a norm-referenced protocol that is sensitive to the learners' age, proficiency level, and the linguistic functions they are required to perform (e.g., writing). We sought to develop a valid language proficiency test that can accurately detect performance variations across distinct language tasks among users of different proficiencies (e.g., L1 and L2 speakers). The validated criterion-referenced language proficiency test would enable the establishment of norm-referenced protocols to accommodate learners of different language backgrounds. For instance, the vocabulary knowledge of a university student who has been using a language as their first and most frequently used language would be better than that of another university student of the same age who only uses the language in the classroom.

1.1.2 Theory of Basic and Higher Language Cognition (Hulstijn, 2015)

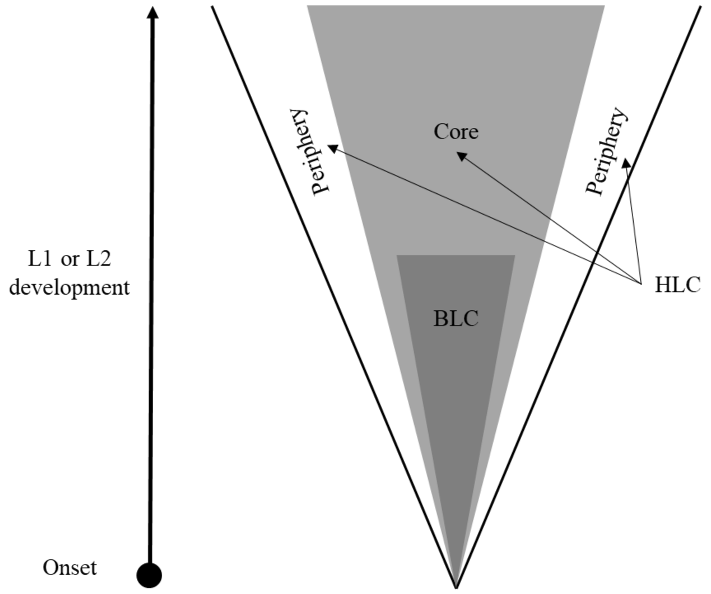
While the Analysis and Control Framework (Bialystok, 2001) untangles the linguistic and metalinguistic demands that characterise various language tasks, the Basic and Higher Language Cognition theory (Hulstijn, 2015; 2019) provides a ground to discuss and compare individual differences in L1 and L2 proficiency (see Figure 1.2) (e.g., Schmid & Yilmaz, 2018; Yi & DeKeyser, 2022). Hulstijn (2015) claims that language proficiency is both the knowledge of language and the ability to access, retrieve and use that knowledge in listening, speaking, reading or writing. Language ability in spoken language (listening and speaking only) that is acquired and shared by all L1 speakers is referred to as basic language cognition, whereas language ability that is not necessarily acquired or shared by all L1 speakers (pertains to both written and spoken languages) is referred to as higher language cognition. Basic language cognition is restricted to frequent lexical items and frequent grammatical structures that is common to all adult L1 speakers, regardless of age, literacy, or educational level. On the other hand, higher language cognition is the domain where differences between L1 speakers can be observed. Language ability in higher language cognition includes low-frequency lexical items or uncommon morpho-syntactic structures, and it applies to both written as well as spoken language. The theory suggests that when a language proficiency test aims at measuring individual differences in L1 proficiency, it should be designed to assess L1 speakers' knowledge of low-frequency lexical items or rare morpho-syntactic structures.

In addition to basic and higher language cognitions, Hulstijn (2015) proposed the core-periphery dimension of language proficiency. Core components of language proficiency, or language ability in the grammatical and pragmatic domains, include knowledge of how to use language forms appropriate to the communicative situation (pragmatic knowledge, sociolinguistic knowledge, knowledge of discourse organization). On the other hand, peripheral components of language proficiency pertain to interactional ability, strategic competence of communicating under adverse conditions (e.g., time constraint) or with limited linguistic knowledge, metalinguistic knowledge (explicit knowledge of grammar), and knowledge of the characteristics of various types of oral and written discourse. For instance, recalling words under time constraints would involve both core (e.g., vocabulary knowledge) and peripheral components of language proficiency (e.g., language strategy under time constraint). Given that the peripheral component is involved and sometimes can be a limiting factor, learners' performance on this task may not completely represent their core language proficiency. If the purpose of a language proficiency test is to assess learners' linguistic knowledge, it is best for the test to only tap into the core components of language proficiency, and minimize the impacts of the peripheral factors.

Taken together, the construct of language proficiency, according to the basic and higher language cognition theory, can be conceptualised in two dimensions: on the one hand, the dimension of basic and higher language cognition, and on the other hand, the dimension of core and peripheral components. Because basic language cognition is assumed to be acquired by all L1 speakers, it is impossible to disentangle the core-periphery components in L1 speakers' basic language cognition (i.e., no variability can be attributed to individual differences induced by the peripheral language proficiency). Therefore, the theory conceptualises both the core and peripheral components of basic language cognition within the core language proficiency (see Figure 1.2). In other words, an L1-L2 difference in basic language cognition (e.g., speeded recognition of high frequency words) is attributed to the difference between core-peripheral components of L2 speakers (e.g., vocabulary knowledge and response strategies under time pressure) and core language proficiency of L1 speakers in basic language cognition (e.g., vocabulary knowledge). Nonetheless, the core and peripheral components of both L1 and L2 speakers are involved in higher language cognition (e.g., speeded recognition of low frequency words), allowing the examination of the relative contribution of each component.

Figure 1. 2

Two dimensions of language proficiency in Basic and Higher Language Cognition theory (Hulstijn, 2015)



Note. Adapted from Hulstijn (2015, pp. 46). BLC: basic language cognition. HLC: higher language cognition. While the core-periphery distinction applies to tasks that assess L2 speakers' BLC and/or HLC, it only applies to L1 speakers language performance with respect to HLC.

Because basic language cognition is assumed to be mastered by all L1 speakers, it can serve as a benchmark for evaluating the language proficiency of L2 speakers. According to the theory, the acquisition of basic language cognition of an L2 is dependent on the type of bilingualism, with age of onset and amount of language exposure and use being taken into consideration (Hulstijn, 2015). Specifically, simultaneous bilinguals who grew up speaking two languages from birth are likely to acquire basic language cognition in both languages if they received sufficient high-quality language input. A similar expectation applies to sequential bilinguals who acquired L2 before school age, or who were raised monolingually and were intensively exposed to L2 between the ages of five and 18 (e.g., through compulsory school education). In contrast, it is unlikely that adult speakers who started learning an L2 after adulthood (e.g., through migration to a country where the L2 is the primary language) will fully master the L2’s basic language cognition. Similar outcomes are expected of L2 speakers who study their L2 as a foreign language in schools, where their L1 is the primary language used in the country.

Based on these expectations, language proficiency tests that aim to capture the proficiency variation in highly proficient bilinguals (e.g., simultaneous and sequential bilinguals) should assess their language ability in higher language cognition (e.g., low frequency words, complex grammatical structure). On the contrary, basic language cognition should be assessed when the focus is on the proficiency of low-to-moderately proficient L2 speakers (e.g., L2 foreign language learners). This thesis aimed to create a language proficiency test that could measure the Malay proficiency of L1 and L2 speakers. Therefore, items that tap into basic and higher language cognition must be carefully sampled in the test development process.

1.2 Language Proficiency Measures

The language proficiency models above proposed that language proficiency test developers should identify and clearly specify the target language component of interest, so that criterion- and norm-referenced protocols can be constructed accordingly with careful consideration of stimulus types (e.g., high or low frequency words) and task demands (e.g., speeded or unspeeded reading). Subsequently, these language proficiency tests can be used in research to inform the distribution of language proficiency among participants or facilitate accurate participant selection for cross-sectional studies that examine the language processing of bilinguals with different language backgrounds.

Measuring language proficiency, however, can be challenging given the complexity of its multidimensional construct (Schoonen, 2011; Treffers-Daller, 2019). When assessing language ability, researchers want to be sure that the test scores capture all relevant components of the target language ability. For example, scores on a multiple-choice reading test are unlikely to capture the rich and complex process of writing a text. Likewise, scores on a cloze test might not fully reflect the vocabulary knowledge of a learner because it involves the use of a combination of vocabulary, grammar, orthographic and semantic knowledge (Read, 2000). The appropriateness of a language proficiency test, therefore, is justified according to the research context, whereby the relationship between the research question, target language proficiency component and how it is measured should be clearly illustrated (Hulstijn, 2010, 2012, 2015; Park et al., 2022; Schmitt et al., 2020; Schoonen, 2011).

Systematic reviews of language proficiency reporting practices (Park et al., 2022; Thomas, 2006; Tremblay, 2011) reveal that both objective (i.e., measures based on the learner's performance on a language task) and subjective language proficiency measures (i.e., measures based on the learner's judgement of they own language level) are commonly used to assess bilinguals' language proficiency. For both approaches, bilinguals' language proficiency is expressed as a score on a scale (e.g., in percentage or seven-point Likert scale) and interpreted based on the construct that the test purports to measure (Hulstijn, 2012). For example, when an objective measure is used (e.g., vocabulary test), a bilingual who receives a higher score is assumed to have a higher level of lexical proficiency than another bilingual who receives a lower score. Likewise, when a subjective measure is used (e.g., self-rated proficiency), a bilingual who rated himself as having "6/*very good*" proficiency in writing is assumed to have higher writing proficiency than a bilingual who has given himself a lower rating.

Therefore, test conceptualisation of a language proficiency measure, including the purpose of the test, target learners, context of testing, aspect, and level of language constructs under measure, is important because it provides guidance on how the scores can be interpreted (Schmitt et al., 2020). These test specifications should be established during test development and validation, and researchers should select the language proficiency measure that matches the experimental context and aim so that the test scores will meaningfully inform the language ability that the experiment aims to investigate. Consequently, this would affect the conclusions made about the relationship between language ability and language processing (Mainz et al., 2017).

To identify the language component that can be used to measure language proficiency (following recommendation from Bialystok, 2001), objective and subjective language proficiency measures that are commonly used in language research are discussed in the following subsections. Some of these measure general language proficiency, whereas others focus on specific language aspects. These measures are developed for a variety of purposes, including curricular decisions in schools, placement purposes in second language classrooms, and independent measures in experimental studies to meet rigorous research standards. Based on how the tests are conceptualised, their utility in research contexts is discussed.

1.2.1 Standardised Language Proficiency Test

When taking language proficiency theories into account, it may seem difficult to create a language proficiency test that accurately measures language learners' general language proficiency because different linguistic and cognitive demands are imposed on the test takers depending on the types of language tasks (e.g., reading or writing), stimuli (e.g., canonical or embedded syntactical structures), and linguistic aspects (e.g., vocabulary or idioms) being tested

(Bialystok, 2001; Hulstijn, 2015). Nevertheless, given the growing number of English-medium universities and the significance of English proficiency to ensure academic success, language proficiency tests have been heavily used to identify prospective students who are able to communicate effectively in English. To assess English learners' language ability in four skill areas, namely reading, writing, listening, and speaking, language proficiency tests such as the Test of English as a Foreign Language Internet-based Test (TOEFL iBT: http://www.ets.org) and the International English Language Testing System (IELTS: https://www.ielts.org) are commonly used (e.g., Ihlenfeldt & Rios, 2023; Ling et al., 2014; Ockey & Gokturk, 2019). These tests are considerably standardised, whereby the instructions, testing conditions, and scoring are designed to follow the same procedures to enable fair testing across various geographical regions (American Educational Research Association et al., 2014).

Standardised tests that aim to measure academic language proficiency use a variety of specific tasks. For instance, while IELTS uses a variety of task formats, including cloze and short-answer questions, TOEFL iBT primarily uses multiple-choice questions. Convergent validity of these standardised test scores have been shown through moderately-strongly correlations with other language task performance (e.g., listening: Nakatsuhara, 2011; Sawaki & Nissan, 2009; oral presentation: Ockey et al., 2015; implicit language knowledge: Erlam, 2006). Because academic success requires a certain level of language proficiency, the test scores are commonly used as a general indicator of the learners' academic language ability for making high-stake academic decisions, such as decisions regarding university admission as well as university students' needs for language support after they have entered the institution (Neumann et al., 2019; Ockey & Gokturk, 2019). However, despite being designed to measure academic language ability that, presumably, could be used to predict academic success, predictive validation studies revealed that the correlation between the scores of the standardised language proficiency test and academic success is only weak to moderate (e.g., Bridgeman et al., 2016; Ginther & Yan, 2018; Neumann et al., 2019; Ihlenfeldt & Rios, 2023). This suggests that language proficiency and academic success are two related but distinct constructs. In addition to language proficiency, academic success could also be affected by several cognitive, social, psychological, and individual factors, including motivation, learning strategies, disciplinary knowledge, and academic acculturation (Fox et al., 2014).

Despite the seemingly holistic assessment of language ability, the comprehensive nature of these standardised language proficiency tests does, however, come at a time and logistics cost. Standardised language proficiency testing is time-consuming and often requires trained assessors to evaluate some assessment components such as the writing and speaking performance. It usually takes around three hours to complete an IELTS or TOEFL test. Moreover, a registration fee of around RM870 (£150) for the IELTS and RM835 (£144) for the TOEFL is needed for each test attempt. Given the tests' duration and costs, they are too long and costly to be used as a standard language proficiency measure in research. Additionally, most standardised language proficiency tests are only available in English (for examples of other standardised English proficiency tests, see Wang et al., 2012 and Wagner, 2020), further limiting testing of understudied languages in research. For example, there is still a lack of standardised language proficiency tests for understudied languages like Malay, despite having 377 million speakers worldwide. Altogether, there remains a call for the creation of practical and valid language proficiency tests for research use, especially for understudied languages. In addition to test validity, these newly developed tests should consider the time needed for implementation and the cost involved for large-scale testing to ensure their practicality in research contexts.

1.2.2 Self-Rated Proficiency

In many languages where practical language proficiency measures are not readily available for research use (e.g., Malay), bilinguals' language proficiency is commonly estimated using self-rated proficiency (e.g., Jalil et al., 2011; Rahman et al., 2018; Y. A. Rusli & Montgomery, 2020). Bilinguals self-evaluate their language proficiency in the four language use areas (i.e., reading, writing, speaking, and listening), typically on a scale of one to seven, with one representing *not at all proficient* in a language and seven being *proficient like a native speaker*. It is quick and easy to collect, often alongside a standardised language history questionnaire (e.g., P. Li et al., 2020; Marian et al., 2007). Its utility as a language proficiency estimate, regardless of target language, is justified by meta-analyses that showed moderate correlations (e.g., $r = .29$ in Zell & Krizan, 2014; $r = .47$ in M. Li & Zhang, 2021) between self-rated proficiency and objectively measured language performances (e.g., vocabulary test scores), and 20.43% of the variance in the objective language performances is accounted for by the ratings (M. Li & Zhang, 2021).

The simplicity of self-rated proficiency, however, is accompanied by some limitations. Although using a standardised language history questionnaire to collect the ratings makes data collection more consistent across studies, the ratings are still subject to variability brought on by individual, between-, and within-group differences (Brysbaert, 2013; Tomoschuk et al., 2019). For instance, L1 speakers may compare their proficiency with other L1 speakers, whereas L2 speakers might refer to the best L2 speaker model they have in mind. Such individual differences in the choice of a proficiency reference could result in unreliable ratings, especially for a heterogenous group (e.g., relatively "noisy" group of participants with a broad range of language proficiency) (Brysbaert, 2013; I. L. Chan & Chang, 2018). In addition, participants of different language combinations (e.g., Spanish-English, Chinese-English) or language background (e.g., heritage speakers or recently immigrated bilinguals) have been found to vary in their accuracy of self-rated proficiency, rendering difficulty in comparing self-rated proficiency across different participant groups (Tomoschuk et al., 2019). Additionally, the stability of self-rated proficiency as a language proficiency estimate is shown to be susceptible to unquantifiable cultural differences (e.g., decision making, personality, degree of exposure to the language in everyday life) (Lemhöfer & Broersma, 2012; Shi, 2011). Thus, measuring language proficiency using self-ratings may not always be as reliable and valid as an objective language proficiency test (L. S. P. Cheng et al., 2021; M. Li & Zhang, 2021; Tomoschuk et al., 2019).

1.3 Vocabulary Knowledge as a Measure of Language Proficiency

The main issues with the standardised language proficiency tests are their cost and length in both administration and scoring. Self-rated proficiency, on the other hand, suffers from various validity issues despite being convenient in terms of time and cost. Given that time and budget are often constricted in research contexts, language proficiency measures that target a specific language ability of interest while also easy to implement are more practical for research settings. To this end, researchers who are looking for a time-efficient and freely available language proficiency measure often use tests that tap into a specific component of language ability.

For instance, researchers investigating syntactical or grammatical knowledge can use elicited imitation tasks as an estimate of language proficiency (e.g., Erlam, 2006). Test takers are expected to be able to verbally produce a sentence after listening to it if they have learned the grammatical features embedded in the stimulus (Rebuschat & Mackey, 2013). However, because sentence repetition involves a range of lexical, phonological, syntactical, and morphosyntactical knowledge, performance on elicited imitation tasks is shown to be more sensitive as a general language proficiency measure than as a discrete linguistic knowledge measure (Yan et al., 2016). Moreover, the systematic review and meta-analysis conducted by Yan and colleagues revealed that the presence of repetition delay, length and grammatical features of the sentence stimuli, as well as the scoring method, would all affect the construct validity of the task. Therefore, the sensitivity of elicited imitation tasks varies greatly across studies because there is no standardisation in how these factors are incorporated in the task design.

On the other hand, vocabulary tests have been commonly used as a proxy for language proficiency by assessing test takers' vocabulary knowledge (Szabo et al., 2021; Treffers-Daller, 2019). This is because vocabulary knowledge is one of the fundamental constructs that underlie language proficiency (Brysbaert et al., 2016; Nation & Beglar, 2007; Qian & Lin, 2020; Schmitt et al., 2015), and it is more straightforward to measure than other aspects of language proficiency (Milton, 2009). For instance, a learner with a score of 20/30 on a vocabulary test is said to know twice as many words as a learner with a score of 10/30. In contrast, it is difficult to assume a learner who received an 8/10 on an essay to have twice the writing ability as a learner who received a 4/10 because language abilities like writing and speaking are more often graded than measured. In such cases, subjectivity of graders and grading criteria also play a role in the grading of the writing skills. Therefore, vocabulary tests have been widely used as an objective estimate of language proficiency in experimental studies that investigate L1 and L2 proficiency for its more quantifiable property compared to other language abilities (e.g., Cop et al., 2015; Fang & Zhang, 2021; Kuperman et al., 2023; Kutlu et al., 2022).

Although assessing vocabulary knowledge might seem more practical than assessing other aspects of language proficiency, it is still not as straightforward as it might seem.

Vocabulary knowledge is a multifaceted unidimensional construct that contains several interrelated but distinct aspects of word knowledge (Durrant et al., 2022; González-Fernández & Schmitt, 2020; Schmitt, 2010, 2014; Webb, 2013). According to Nation (2013, 2020, 2022), mastery of nine aspects of word knowledge is required to achieve lexical proficiency, including knowledge of various word forms, meanings, and uses of a word (see Table 1.1). Each aspect can be further divided into receptive and productive knowledge. The receptive/productive conceptualisation entails how various word knowledge are used for communicative purpose in real life. Receptive knowledge refers to the skills needed to recognise 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 connection (e.g., recognising “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, and other aspects of word knowledge (e.g., constraint on 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.

Table 1. 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 spelled?
	Word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express the meaning?
Meaning	Form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	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).

It is, however, difficult to truly measure distinct word knowledge aspect in isolation based on the skill-based receptive/productive definitions (Schmitt, 2010). Alternatively, researchers who seek to measure aspect-specific word knowledge commonly assess recognition and recall of word knowledge aspects 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 knowledge needed to recognise and select target from an array of choices, while a word recall task assesses knowledge needed for target retrieval after certain cues such as a figure illustration or the word meaning is presented. Overall, word recognition has been shown to precede the acquisition of word recall (González-Fernández & Schmitt, 2020). For instance, using the recognition and recall tasks to assess form-meaning knowledge (see Table 1.2 for the task formats used to examine four levels of form-meaning knowledge), previous studies (Laufer & Aviad-Levitzky, 2017; Aviad-Levitzky et al., 2019; Laufer & Goldstein, 2004; Schmitt, 2010) revealed that the mastery levels of form-meaning knowledge are implicationally scaled, in which 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 a word's meaning, would depend on the form-meaning knowledge that was acquired earlier, such as 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.

Table 1. 2

Levels of mastery in form-meaning link (Laufer & Goldstein, 2004; Schmitt, 2010)

Aspect of word knowledge		Test format	Task	Example
Given	Tested			
Form	Meaning	Recognition	Select definition or translation in L1	cat: *a. kucing* *b. anjing* *c. tikus* *d. burung*
	Meaning	Recall	Supply definition or translation in L1	cat: *k*__________
Meaning	Form	Recognition	Select word in L2	*kucing:* a. cat b. dog c. mouse d. bird
	Form	Recall	Supply word in L2	*kucing:* c__________

Note. Adapted from Schmitt (2010). The degree of form-meaning knowledge is labeled by matching the aspect of word knowledge being tested with the relevant test format, e.g., meaning recognition.

Taken together, vocabulary testing can be complicated because of the multifaceted nature of the vocabulary knowledge construct. To maximise their usefulness in research, the types and components of vocabulary knowledge measured by vocabulary tests should be carefully considered (Schmitt et al., 2020). In the next section, the facets of vocabulary knowledge and some commonly used vocabulary tests are discussed.

1.3.1 Depth and Breadth of Vocabulary Knowledge

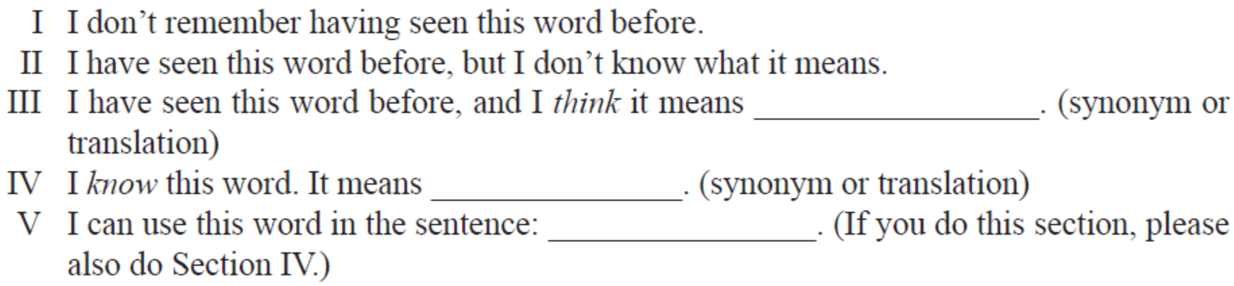
There is currently no consensus about how vocabulary knowledge can be accurately measured in view of its multifaceted and interrelated nature (Durrant et al., 2022; González-

Fernández & Schmitt, 2020; Schmitt, 2010, 2014; Webb, 2013). In general, vocabulary knowledge can be measured in two ways: depth and breadth (Anderson & Freebody, 1981; Durrant et al., 2022; Schmitt, 2014; Webb, 2013). Depth of vocabulary knowledge refers to the quality of vocabulary knowledge. It is conceptualised as the overall degree of knowledge of all the word knowledge aspects involved (e.g., knowledge of collocation: how words should be used together, and word association: how different words may be used interchangeably) (Nation, 2013). The depth of vocabulary knowledge has been defined and conceptualised in the literature in many ways, and accordingly, different approaches have been used to measure vocabulary depth (Yanagisawa & Webb, 2020). Some of the key approaches will be discussed in the following paragraphs.

Depth of vocabulary knowledge can be viewed as a spectrum from no word knowledge to fully developed word knowledge. Along with this conceptualisation, the Vocabulary Knowledge Scale (Paribakht & Wesche, 1993) uses a combination of self-assessment and productive items to capture the developmental stage of word knowledge. Test takers are asked to indicate their degree of knowledge of each target word using a five-stage scale, and different scores are assigned based on the stage level chosen for each vocabulary item (see Figure 1.3).

Figure 1. 3

The five-stage scale in the Vocabulary Knowledge Scale (Paribakht & Wesche, 1993)



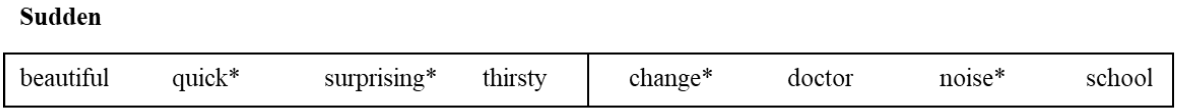
I I don't remember having seen this word before.
II I have seen this word before, but I don't know what it means.
III I have seen this word before, and I *think* it means _________________. (synonym or translation)
IV I *know* this word. It means _______________. (synonym or translation)
V I can use this word in the sentence: ________________. (If you do this section, please also do Section IV.)

Note. Adapted from Paribakht and Wesche (1993).

On the other hand, knowledge of word association is tested when vocabulary depth is conceptualised as a lexical network in which learners' vocabulary is linked in their mental lexicon. The Word Associates Test (Read, 1993, 1998) is one of the most widely used tests to assess the connections between words within mental lexicon via multiple aspects of vocabulary knowledge, including synonymy, polysemy, and collocations. In this test, test takers are required to identify four words out of eight choices that are associated with the target words that either have a paradigmatic (have a related meaning) or syntagmatic (appear together in context, i.e., collocate) relationship with the target word (see Figure 1.4).

Figure 1. 4

An example from the Word Associates Test (Read, 1993, 1998)



Sudden

beautiful	quick*	surprising*	thirsty	change*	doctor	noise*	school

Note. Adapted from Read (1998). The correct answers are marked with asterisks.

In addition, the depth of vocabulary knowledge can also be conceptualised by distinct components involved in knowing a word. This corresponds to the word knowledge aspects proposed by Nation (2013; see Table 1.1). Research adopting this conceptualisation has either measured one or several aspects of word knowledge. For instance, Webb (2005) created 10 tests that tap into five aspects of word knowledge (i.e., written form, form and meaning, association, collocation, and grammatical functions), each can be assessed with receptive or productive formats. On the contrary, Nguyen and Webb (2017) focused only on one aspect of word knowledge (i.e., collocation) and developed a multiple-choice test in which test takers were required to choose the word that co-occurred most frequently with the target word. Among the three types of conceptualisations, this approach is the more common approach because measurement of distinct aspects allows more direct and transparent test score interpretation, such that it is clear what type of word knowledge the test score represents (Webb, 2013; Yanagisawa & Webb, 2020).

Nevertheless, a test format must be determined in order to assess the criterion-referenced performances of language learners (Bialystok, 2001). Unfortunately, there is currently no consensus in the literature about the best approach to operationalise and measure the depth of vocabulary knowledge (Durrant et al., 2022; González-Fernández & Schmitt, 2020; Schmitt, 2014; Webb, 2013). Furthermore, some tasks that measure vocabulary breadth, in practice, may also involve measuring depth of vocabulary knowledge incidentally (Schmitt, 2014). For instance, counting the number of words known by testing learners' knowledge of form-meaning connections (see Table 1.1), either by recognising or supplying the word form based on its meaning, involves the testing of form-meaning knowledge to a certain degree of depth. This extent to which a single word knowledge aspect is known is referred to as the strength of vocabulary knowledge, and it is different from the depth of vocabulary knowledge (Nation & Webb, 2011; Webb, 2013). For example, the strength of form-meaning knowledge can range from receptive recognition and recall, which include identification of a word form or meaning when given its counterpart, to productive recognition and recall, where one can produce the word form or meaning when given its counterpart (Laufer & Goldstein, 2004; Schmitt, 2010). In other words, the strength of knowledge is used to describe the degree of knowledge in one aspect of word knowledge, whereas the depth of knowledge refers to the quality of multiple aspects of word knowledge. This further highlights the complexity involved in the assessment of vocabulary depth, because testing only one aspect of word knowledge (e.g., form-meaning connection) may not be sufficient to represent test takers' comprehensive depth of vocabulary knowledge (Nation & Webb, 2011; Schmitt, 2014).

Because of the complexity of measuring depth of vocabulary knowledge, most vocabulary tests designed for research have focused on the breadth of vocabulary knowledge, or the number of words known by a person (Schmitt, 2014). These tests are commonly known as vocabulary size tests. By tapping into the dimension of form-meaning knowledge, scores of vocabulary size tests are used to estimate performance in various language tasks. For instance, Nation (2006) showed that knowledge of at least 8,000 word-families (i.e., groups of words each share a common root word, such as "help", "helpful" and "helpless") is needed for language learners to perform various language tasks fluently (e.g., reading newspaper, watching movie), suggesting the importance of vocabulary knowledge establishment for other language abilities. Furthermore, vocabulary size has strong correlations with various aspects of word knowledge (e.g., collocations, multiple meanings) (González-Fernández & Schmitt, 2020) and word processing (e.g., listening comprehension) (Andringa et al., 2012; Rodríguez-Aranda &

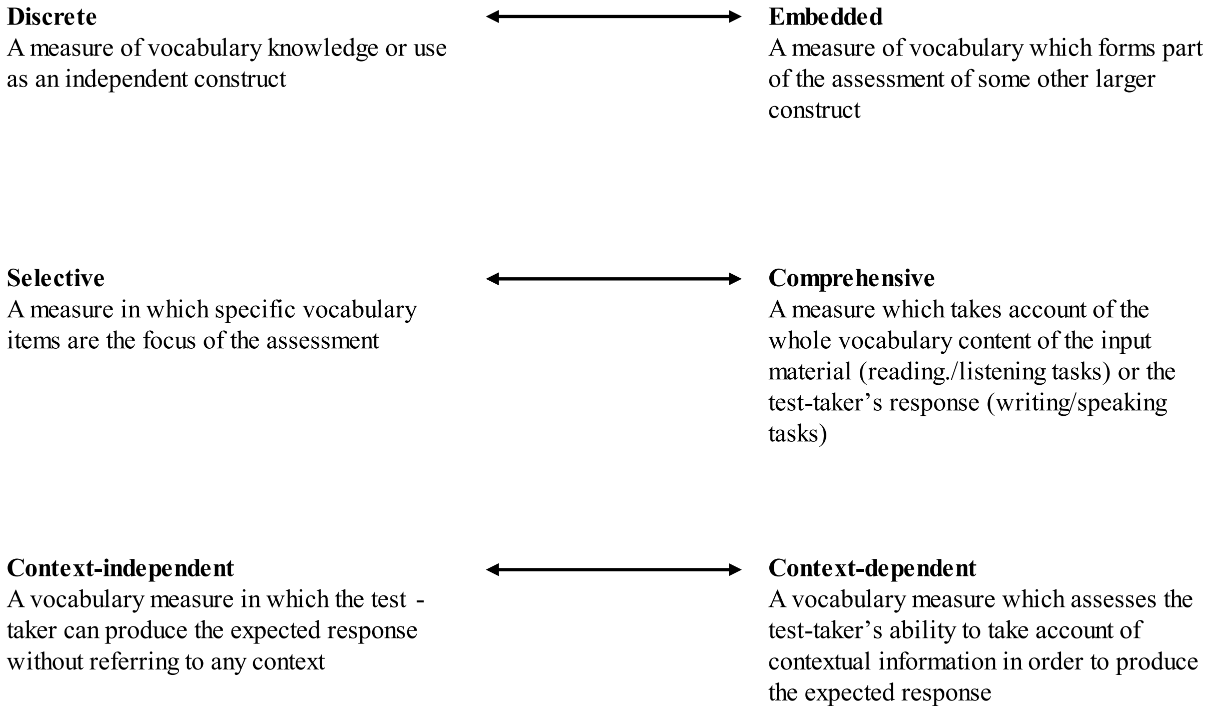
Jakobsen, 2011; M. J. Yap et al., 2012). Taken together, these findings support the use of a vocabulary size test as a language proficiency estimate in bilingual research to account for individual differences in terms of language proficiency or ability. In the following subsection, the framework by Read (2000) is reviewed to illustrate the different dimensions of vocabulary assessment and how they can affect test score interpretation.

1.3.2 Dimensions of Vocabulary Assessment

Read (2000) proposed three dimensions in which vocabulary assessment can be designed to measure vocabulary knowledge (see Figure 1.5). The discrete – embedded dimension focuses on the target construct that a vocabulary test measures, considering the purpose and score interpretation of the test. A discrete vocabulary test assesses vocabulary knowledge as a distinct construct, separated from other components of language ability. This way, it allows interpretation of the scores as a measure of learners' vocabulary knowledge. In contrast, an embedded vocabulary test assesses vocabulary knowledge within a larger construct. For example, knowledge of words can be assessed via comprehension questions following a written passage, in which the learners' understanding of particular words is tested. In these tasks, it is more difficult to parse vocabulary knowledge from other language ability (e.g., grammatical knowledge) based on the embedded vocabulary test scores alone. Instead, it forms a part of a measure for a larger language construct, such as reading comprehension.

Figure 1. 5

Dimensions of vocabulary assessment



Discrete
A measure of vocabulary knowledge or use as an independent construct

Embedded
A measure of vocabulary which forms part of the assessment of some other larger construct

Selective
A measure in which specific vocabulary items are the focus of the assessment

Comprehensive
A measure which takes account of the whole vocabulary content of the input material (reading./listening tasks) or the test-taker's response (writing/speaking tasks)

Context-independent
A vocabulary measure in which the test - taker can produce the expected response without referring to any context

Context-dependent
A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response

Note. Adapted from Read (2000, pp. 9).

The selective – comprehensive dimension, on the other hand, refers to the breadth of vocabulary knowledge being assessed (Read, 2000). A comprehensive vocabulary test considers the vocabulary content in a spoken or written context, as opposed to a selective vocabulary test that evaluates learners' vocabulary knowledge based on a predetermined set of target words. For instance, a selective vocabulary test assesses the production of names for a set of target pictures, whereas a comprehensive vocabulary test rates the overall quality of vocabulary used in a spoken language context.

Vocabulary tests can measure learners' ability to understand or use target words in either context-independent or context-dependent manner (Read, 2000). Context-independent vocabulary tests do not require contextual information (e.g., a meaningful sentence) to induce appropriate responses to the test items, whereas context-dependent vocabulary tests require some understanding of the context in order to provide correct responses (e.g., S. Zhang & X. Zhang, 2022).

Taken together, Read’s (2000) framework serves as guidance for vocabulary test development and implementation (e.g., Amenta et al., 2020; Masrai, 2022). It is advocated that the vocabulary test format and context would affect how the test scores could be interpreted, however there is no expectation that one type of test would be better than the other. For instance, scores from a discrete, selective, and context-independent vocabulary test can be interpreted as a measurement of construct-specific ability (i.e., the target aspect of vocabulary knowledge), because they are based solely on the test taker's knowledge of a predetermined set of words without reference to any context. In the following subsection, some commonly used vocabulary tests and the dimensions in which the tests are designed are discussed in relation to the Read’s (2000) framework.

1.3.3 Current Assessments of Vocabulary Breadth

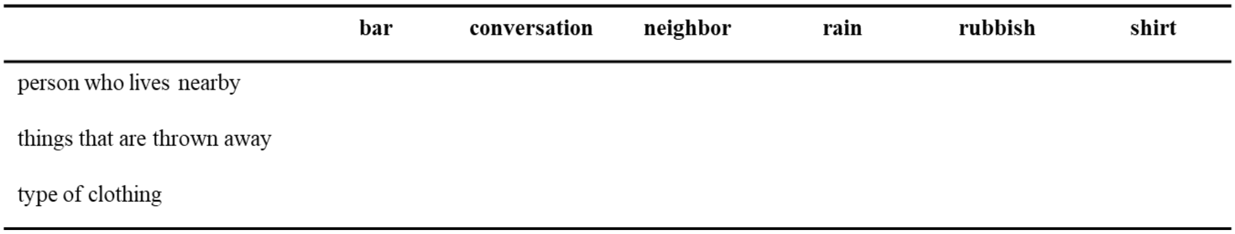
Vocabulary size tests measure the extent to which a list of words is known by a test taker. These tests come in a variety of formats; some used selected-response items, such as multiple matching (e.g., Webb et al., 2017), multiple-choice (e.g., Nation & Beglar, 2007), and yes/no items (e.g., S. T. Lee et al., 2023; Lemhöfer & Broersma, 2012; Masrai, 2022), whereas others employed constructed-response items, such as translation production that involves form and meaning recall (e.g., McLean et al., 2020). Because different task formats elicit different levels of form-meaning knowledge, with some being cognitively more demanding than the others, the inferences that can be drawn from test scores depend on the test conceptualisation, specifically the purpose (e.g., reading comprehension) and dimensions of the tests (e.g., discrete, selective, and context-independent), as well as the choice of test format (e.g., yes/no lexical decision).

1.3.3.1 Vocabulary Levels Test

The Updated Vocabulary Levels Test (Webb et al., 2017; see Nation, 1983 and Schmitt et al., 2001 for the earlier versions) is a discrete, selective, and context-independent vocabulary test developed to assess learners' form recognition at the first five 1000-word frequency levels from the British National Corpus/Corpus of Contemporary American English (Nation, 2012a). The test uses a form-recognition matching format, in which three word-meanings and six word-forms (three targets and three foils) are presented in a cluster (see Figure 1.6 for an example of test items from the Updated Vocabulary Levels Test). Test takers' task is to select the word form that matches with each of the three meanings provided. Because the test is designed to tap into the initial mastery level of form-meaning knowledge, the foils in each cluster have very different meanings so that they allow for partial knowledge, or words that are only partially understood by test takers (e.g., knowing that pomelo is a type of fruit). For instance, given the choices with no overlapping in meanings (see Figure 1.6), test takers who have the impression that "neighbour" refers to a person would be able to make the correct match even though they do not know the full meaning of the word that refers to "someone living nearby". The test score (out of 30 items) for each frequency level serves as a measure for the mastery of L2 vocabulary knowledge at specific frequency levels. Therefore, the Updated Vocabulary Levels Test can be used to provide a vocabulary profile for language learners, which is particularly useful to advise the most appropriate frequency level for the learners' vocabulary learning in research (e.g., Dang, 2020; Ha, 2021).

Figure 1. 6

An example of the form recognition matching format used in the Updated Vocabulary Levels Test (Webb et al., 2017)



	bar	conversation	neighbor	rain	rubbish	shirt
person who lives nearby						
things that are thrown away						
type of clothing						

1.3.3.2 Vocabulary Size Test

In addition to the Updated Vocabulary Size Test (Webb et al., 2017), the Vocabulary Size Test (Nation & Beglar, 2007) is another prominent vocabulary test in English language research. It is a discrete, selective, and context-independent meaning-recognition test intended to provide an estimate of English L1 and L2 speakers' overall receptive vocabulary size. The test employs a multiple-choice format, with 140 items that examine knowledge of English words from a wide word frequency range (from 1000 to 14000 frequency level, with 10 items selected at each 1000 frequency level). The 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 meaning that matches with the target word presented.

Similar to the Vocabulary Levels Test, test takers could score in the Vocabulary Size Test with just partial knowledge of the vocabulary items because the test foils do not share core elements of the target word meaning. Difficulty level between the target words and their foils is matched by making sure that the foils share the same word class and frequency band as the target words (Nation & Beglar, 2007). To ensure discrete testing of vocabulary knowledge (i.e., to only test vocabulary knowledge and not beyond), meaning choices of the test are usually written in easier language than the target word, in which the words used in the choices are, to the greatest extent, of higher frequency than the item being defined. Taken together, the Vocabulary Size Test measures learners' overall vocabulary size by taking into account both partial and complete form-meaning knowledge through meaning recognition. The test has been used as a vocabulary size measure in language research to investigate its relationship with other language performances (e.g., meaning recall: Stoeckel et al., 2019; word association: Janebi Enayat & Amirian, 2020).

1.3.3.3 Lexical Test for Advanced Learners of English

In the field of psycholinguistics, the Lexical Test for Advanced Learners of English (LexTALE) developed by Lemhöfer and Broersma (2012) has been widely used to measure English proficiency of advanced learners of English. This discrete, selective, and context-independent vocabulary test takes an unspeeded yes/no lexical decision format. It contains a total of 60 test items (40 words and 20 nonwords) with the ratio of words and nonwords being 2:1 and takes about 5 minutes to complete. In the task, test takers indicate if letter strings are existing English words by responding "yes" or "no". LexTALE is freely available in the form of paper-and-pencil and online formats. Previous studies have demonstrated validity of LexTALE by showing its ability to explain language performance measured by other language tasks such as the lexical decision and visual word recognition tasks (e.g., Diependaele et al., 2013; Lemhöfer & Broersma, 2012; Wen & van Heuven, 2017b). Furthermore, LexTALE has been used to assess the language proficiency of English monolingual speakers, and it has been shown to capture a substantial variability among highly proficient speakers (Diependaele et al., 2013). Therefore, objective language proficiency measures like LexTALE are encouraged to be used as standard language proficiency measures in bilingual research to allow generalisation and comparison of research findings across studies (Diependaele et al., 2013; Lemhöfer & Broersma, 2012; H. Zhang et al., 2020).

It is, however, important to note that LexTALE is specifically designed and validated for moderate-highly proficient English L2 speakers. By focusing on L2 speakers within this proficiency range, the test items were carefully selected to avoid too many items being unknown to the test takers, which may inflate the guessing/false alarm rate (e.g., test takers giving up because the test is too difficult or responding with 'yes' too frequently). In addition to English, it also has parallel versions in Dutch and German, designed with their difficulty level being matched as closely as possible to allow cross-linguistic comparisons (Lemhöfer & Broersma, 2012). Inspired by LexTALE and its extensions, other researchers have developed similar lexical tests to estimate proficiency of other languages. To date, lextale extensions are available for French (LEXTALE-FR: Brysbaert, 2013), Spanish (Lextale-Esp: Izura et al., 2014), Chinese (LEXTALE_CH: I. L. Chan & Chang, 2018; LexCHI: Wen et al., 2023), Italian (LexITA: Amenta et al., 2020), Portuguese (LextPT: Zhou & Li, 2022) and Finnish (Lexize: Salmela et al., 2021). These lextale extensions were not matched against LexTALE in terms of the word stimuli used and their difficulty level. Instead, they were designed to measure vocabulary size of speakers from a wider language proficiency range (i.e., L1 and L2 speakers). In these tests, more items were included, and overall difficulty level was increased to improve their reliability and suitability to measure language proficiency of both L1 and L2 speakers (Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014).

LexTALE and its extensions have been widely used as a lexical proficiency measure in empirical studies to match participants’ proficiency across different groups as well as discriminate between speakers of different proficiency levels. For instance, Puig-Mayenco et al. (2023) showed that LexTALE has been used in 551 studies in the past decade as a measure of lexical proficiency for speakers of different L1 (e.g., Dutch, Korean, Spanish). In addition, lextale extensions, such as LEXTALE-FR (Brysbaert, 2013) and Lextale-Esp (Izura et al., 2014), have also been used as a lexical proficiency measure in non-English L2 experiments (e.g., Dujardin et al., 2022; Sarrett et al., 2022).

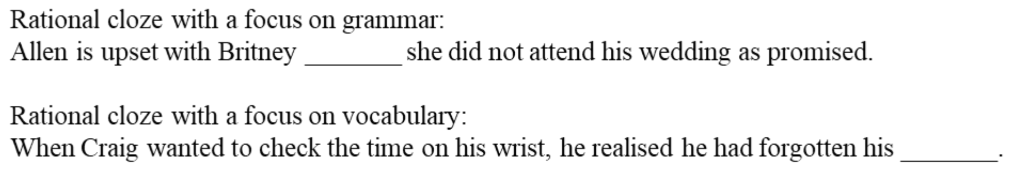
1.3.3.4 Cloze Tests

In contrast to discrete and context-independent vocabulary tests, cloze tests serve as a distinct test type that assesses vocabulary knowledge in a context-dependent and embedded format (Read, 2000). Cloze tests contextualise vocabulary test items while maintaining some degree of control over the items being tested (c.f. complete control of target stimuli in selective tests, e.g., LexTALE). Using a fill-in-the-blank format, cloze tests require test takers to fill predetermined gaps embedded in written texts (e.g., a long paragraph). A variety of cloze formats have been used in language testing, each necessitates the use of vocabulary knowledge to different degrees (Chapelle, 1994; Eckes & Grotjahn, 2006; Harsch & Hartig, 2016; Read, 2000; Singleton, 1999).

In a standard cloze test, words are deleted from written text in a predetermined ratio (e.g., every sixth word). The fixed-ratio deletion removes a variety of word types (e.g., nouns, verbs, and prepositions) from a sentence, making assessment of general word knowledge possible (Read, 2000; Schmitt, 2010). However, when words are deleted at a fixed ratio, the word types and information carried by the deleted words become unpredictable (e.g., when every seventh word is deleted, the deleted words can be a noun in the first sentence but an adverb in the second), affecting the consistency of the linguistic component being measured. On the other hand, the rational cloze deletes words based on predetermined linguistic criteria (e.g., grammar, content word) to test specific linguistic knowledge, providing a more deliberate method for deleting words from the text (see Figure 1.7 for examples of different target linguistic criteria).

Figure 1. 7

Example of rational cloze questions



Rational cloze with a focus on grammar:
Allen is upset with Britney _______ she did not attend his wedding as promised.

Rational cloze with a focus on vocabulary:
When Craig wanted to check the time on his wrist, he realised he had forgotten his _______.

Fixed-ratio and rational cloze formats can be further modified into multiple-choice cloze tests. Instead of producing the deleted words in written form, a multiple-choice cloze test turns the cloze questions into multiple-choice items, in which learners are required to select the appropriate option to fill in the blank. In other words, unlike fixed-ratio and rational cloze tests, the multiple-choice cloze tests assess vocabulary knowledge through language comprehension rather than production.

The C-tests, on the other hand, delete the second half of every second word in a written text (Read, 2000; Cromheecke & Brysbaert, 2022). For example, the test takers are required to fill in the gaps in the following text: “This is an example of C-tests. Please fi__ in th__ blanks wi__ missing bi__ of infor______.” Because correct responses not only require vocabulary knowledge but also grammatical and morphological knowledge, it is widely regarded as a reliable and valid measure of general language proficiency because it taps into a wide range of language skills and linguistic knowledge (e.g., Eckes & Grotjahn, 2006; Gellert & Elbro, 2013;

Harsch & Hartig, 2016; Klein-Braley, 1985). However, for researchers who are particularly interested in vocabulary knowledge, the embedded nature of the C-tests is less useful as a distinct vocabulary measure because answering the items requires linguistic knowledge beyond vocabulary knowledge (Eckes & Grotjahn, 2006; Gellert & Elbro, 2013; Read, 2000).

1.3.3.5 Translation Task

Translation serves as a distinct, selective, and context-independent vocabulary task to assess vocabulary knowledge. In contrast to meaning recognition formats (see Sections 1.3 and 1.3.1) that account for partial knowledge, translation tasks assess the precise recall or production of form-meaning knowledge without cues (see Table 1.2; Laufer & Goldstein, 2004; Schmitt, 2010; Stewart et al., 2023). Forward translation (L1 to L2 translation) tests learners' form recall knowledge by asking them to supply a word form in their L2 based on the meaning of L1 words that are assumed to be fully known. On the other hand, meaning recall knowledge is tested when learners perform backward translation (L2 to L1 translation), in which they are required to provide an equivalent L1 word form to demonstrate knowledge of meaning in the L2 words.

Previous research suggests that recall formats are arguably better vocabulary knowledge measures than meaning-recognition formats in testing form-meaning knowledge of the same words because they assess form-meaning knowledge at a higher level (McLean et al., 2020; Stewart et al., 2023; Stoeckel et al., 2019). Knowledge of meaning and form recall usually develop after that of recognition, thus implying greater mastery of form-meaning knowledge (McLean et al., 2020; Laufer & Goldstein, 2004; González-Fernández & Schmitt, 2020). Put differently, when one demonstrates recall knowledge, he is more than likely to also have mastered the lower forms of form-meaning knowledge (i.e., form and/or meaning recognition).

1.3.4 Language Proficiency Measure for Research Use

In summary, there is a great variability in how language proficiency is operationalised and measured in language research. In contrast to high-stake decision makers at higher education who are interested to know about the learners' overall language proficiency, language researchers are more interested in using practical low-stake language tests that can provide valid and reliable estimates of learners' proficiency on specific language aspect of interest (e.g., vocabulary knowledge). Therefore, while standardised language proficiency tests (e.g., IELTS and TOEFL) are useful in academic settings to guide important decisions, given their length and cost involved, they may not be the best choice for research use.

Vocabulary size tests, on the other hand, are a popular choice in research because their scores correlate well with other language performances (e.g., Diependaele et al., 2013; Lemhöfer & Broersma, 2012; Wen & van Heuven, 2017b; S. Zhang & X. Zhang, 2022). Aside from their validity as a measure of language proficiency, they are often freely available and time-efficient, making them popular and ecologically valid language proficiency measures for research use. However, different formats of vocabulary size tests place different cognitive and linguistic demands on the test takers, subsequently leaving an impact on test score interpretation (Laufer & Aviad-Levitzky, 2017; Schmitt, 2010; Schmitt et al., 2020; Stewart et al., 2023). As a result, the validity of these tests is likely to be dependent on the proficiency level of the target participants (e.g., L1 vs. L2 speakers) and the criterion to be predicted (Cromheecke & Brysbaert, 2022).

For instance, although recall tests measure higher levels of form-meaning knowledge, it does not undermine the value of recognition tests as vocabulary knowledge measures. There is a distinction in the types of vocabulary assessed by the two test formats. Recall tests are useful to assess sight vocabulary, or words that learners comprehend upon seeing them in isolation or without a context (Laufer & Aviad-Levitzky, 2017). Sight vocabulary can be precisely identified and understood outside of context or without cues; therefore, it can be tested in recall tests with no cues provided. However, recall tests might not be able to capture words that learners only partially understand (i.e., partial knowledge vocabulary; Laufer & Aviad-Levitzky, 2017; see example of "neighbour" in Section 1.3.3.1) and words that are known but whose meaning cannot be retrieved at a particular moment (i.e., cued recall vocabulary; Laufer & Aviad-Levitzky, 2017). Particularly, recognition tests are useful if the purpose of testing is to capture knowledge of these two types of vocabulary (in addition to sight vocabulary). Multiple-choice options in recognition tests, such as the Vocabulary Levels Test and the Vocabulary Size Test have no overlap in the core meaning of target words. This strategic design could be used to trigger learners' memory, thus allows the assessment of cued recall vocabulary while accounting for partial vocabulary knowledge (Laufer & Aviad-Levitzky, 2017). Taken together, the degree to which the different test formats assess form-meaning knowledge clearly depicts their distinct value in testing vocabulary knowledge. The appropriateness of the test formats in measuring vocabulary knowledge should therefore be determined by the intended purpose of the vocabulary testing (Schmitt et al., 2020; Webb, 2021).

Taken together, recall tasks (e.g., bidirectional translation tasks) might have the highest predicting power for reading comprehension (e.g., McLean et al., 2020; S. Zhang & X. Zhang, 2022), whereas recognition tasks might be more useful in predicting response time in word recognition (e.g., Diependaele et al., 2013). Therefore, appropriateness of language proficiency measure for research use is determined by the convergence between the research questions and the conceptualisation of the chosen vocabulary test. The purpose of the vocabulary test, target learners, context of testing, aspect of vocabulary knowledge under measure (e.g., form-meaning connections) and its mastery level (e.g., form recognition) should all be considered to justify the utility of a vocabulary test and its score interpretation (Schmitt et al., 2020).

Despite the significance of these considerations in test development, it is important to note that the majority of existing language proficiency measures are mainly available in English. The insights obtained from the theories of bilingual language acquisition and testing (Bialystok, 2001; Hulstijn, 2015), along with the utility of various vocabulary test formats in assessing bilingual speakers of languages beyond European languages, are rarely applied to other understudied languages, including Malay. The scarcity of reliable language proficiency tests for these understudied languages not only limits research opportunities but also hinders the rigour of research conducted in these languages. Therefore, to enhance language testing for research in these languages, there is a need to obtain a deeper understanding of the language proficiency measures that are appropriate for the language users. As a starting point, this thesis focuses on Malay, an understudied Austronesian language, and investigates the utility of various language proficiency measures in measuring the proficiency of Malay-speaking bilinguals. The following section delves into the unique multilingual landscape of Malaysia and discusses the linguistic characteristics of the Malay language. It also discusses the challenges of language proficiency testing in the country, particularly in the absence of appropriate language assessment tools for this understudied language.

1.4 Bi-/Multilingualism in Malaysia

Malaysia, a country located in Southeast Asia, stands as a multilingual and multiracial nation. With Malays being the majority race in Malaysia (59.1%), followed by Chinese (23.0%) and Indian (6.7%) (Department of Statistics Malaysia, 2021), the Malay language, or Bahasa

Malaysia (i.e., the mother tongue of Malays), is used as the country's official language. Vernacular languages such as Mandarin are used by the native speakers within the local communities alongside the official language. In addition to Malay and vernacular languages, the Malaysian government promotes the use of English, such that bilingual education of Malay and English languages is emphasised in the Malaysian National Education Blueprint 2013–2025 (Kementerian Pendidikan Malaysia, 2013). Specifically, Malay serves as the primary language of instruction (except for the Mandarin and Tamil vernacular schools) and a compulsory subject in all schools (S. H. Chan & Abdullah, 2015; Mostafa, 2016). More importantly, English has been taught as a mandatory second language in all schools (S. H. Chan & Abdullah, 2015; Mostafa, 2016). On an additional note, inclusion of subjects for vernacular languages like Mandarin and Tamil is compulsory in the respective vernacular schools, but optional in national schools. As a result, most Malaysians are bi-/multilinguals who are proficient in at least two languages (i.e., Malay and English); for the Malay L1 speakers, Malay as the national language is acquired as their first language, and English is learned as a compulsory second language. For the other races, the vernacular language (e.g., Mandarin) is commonly acquired as the first language, with Malay and English being learned as the second or third languages (Mahmud & Salehuddin, 2023).

1.4.1 The Malay Language

Being a language from the Austronesian language family, Malay is commonly investigated in psycholinguistic research for cross-linguistic comparisons with English (e.g., Y. A. Rusli & Montgomery, 2020). Both Malay and English share the same 26 letters, but the former has shallower orthography depth, simpler syllable structures, and more transparent affixation compared to the latter (M. J. Yap et al., 2010). Furthermore, Malay possesses a more complex morphological system, where words can be formed via rule-based affixation (M. J. Yap et al., 2010). For instance, a noun (e.g., “*penulis*/author”) can be formed by adding a noun prefix “*peN-*” to a verb “*tulis*/write”. In a similar way, an adjective (e.g., “*bertulis*/having writing”) can be formed by adding a verb prefix “*ber-*” to the same word. In view of these morphological differences, Malay words have more syllables and a wider range in word length compared to English words (L. C. Lee et al., 2007). Taken together, cross-linguistic research involving Malay and English could generate important insights regarding the effects of different language-specific characteristics (e.g., morphological complexity, orthography depth) on bilingual language processing.

Psycholinguistic studies conducted in Malaysia often use Malay for cross-linguistic comparisons with English because of the rather unique population of bilingual Malay-English speakers in Malaysia (e.g., Rahman et al., 2018; Y. A. Rusli & Montgomery, 2020). Many Malaysians can be considered early Malay-English bilinguals because both languages were taught and acquired when they first started school (Jin et al., 2013). Furthermore, both Malay and English are commonly used in various daily contexts. This enables a good proportion of Malaysians to become highly proficient in both languages and therefore interesting for exploration of various research questions on bilingualism (e.g., see Duñabeitia et al., 2010; Hulstijn, 2015). Nevertheless, despite their early exposure to the two languages, their language proficiency might still differ due to differences in their language learning ability and degree of language usage. Due to the bilingual educational policy in the country (which is further complicated by the vernacular school system), it is challenging, if not impossible, to estimate the differences in the bilinguals’ language proficiency based solely on their order of acquisition or language exposure. Therefore, a quick objective test of Malay proficiency would be very useful for research use in this population.

1.4.1 Measuring Malay Language Proficiency

To the best of my knowledge, there is yet no freely available Malay proficiency measure that is designed for research use. Studies that involved Malay-speaking bilinguals have so far either assumed "native-like" proficiency of Malay L1 speakers (e.g., L. W. Lee & Low, 2014; N. T. Yap et al., 2017), or used self-ratings to estimate the speakers' language proficiency (e.g., Jalil et al., 2011; Rahman et al., 2018; Y. A. Rusli & Montgomery, 2020). However, the assumption of "native-like" proficiency is not always reliable because even the language proficiency of L1 speakers (e.g., vocabulary size) could vary substantially according to the speakers' language experience (e.g., whether a person reads) (Brysbaert et al., 2016; Hulstijn, 2015, 2019). Furthermore, language proficiency estimated by self-ratings can be affected by individual and group differences (see discussion under 1.2.2). Thus, assuming "native-like" proficiency or using self-ratings to measure language proficiency may not always be reliable and valid as a language proficiency estimate in bilingual research (L. S. P. Cheng et al., 2021; M. Li & Zhang, 2021; Tomoschuk et al., 2019). Thus, there is a need for a valid language proficiency test that could reliably quantify the language proficiency of Malay speakers with different proficiency levels.

This thesis sought to enhance language proficiency testing in Malaysia. Due to the lack of research resources available in the language, three empirical studies were conducted to identify appropriate language proficiency measures to assess the proficiency of Malay L1 and L2 speakers. In particular, vocabulary size test was chosen for the thesis because vocabulary knowledge has been shown to be the foundation for other language abilities (Nation, 2006), and vocabulary size correlated strongly with various aspects of word knowledge (Andringa et al., 2012; González-Fernández & Schmitt, 2020; Rodríguez-Aranda & Jakobsen, 2011; M. J. Yap et al., 2012). Therefore, various vocabulary test formats that were commonly used in research, including translation tasks, yes/no vocabulary tests, cloze tests, and form-meaning vocabulary tests, were examined in this thesis. The following section provides an overview of this thesis.

1.5 Thesis Outline

Given the lack of a freely available language proficiency measure in Malay, the aim for this thesis is threefold. Given the lack of language testing resources in Malay, translation norms for Malay and English were first collected to investigate translation performance of the bilinguals. Following the development of translation norms, a Malay language proficiency test that is suitable for research utility was created and subsequently validated using an array of language tasks that serve as external criterion measures. Data collected from the translation norms served as a foundation for stimuli selection of these criterion measures.

The following chapter presents the Malay-English and English-Malay translation norms as a database to facilitate stimulus selection for Malay-English cross-linguistic research. The prevalence of translation ambiguity between Malay and English is expected to be higher than that of previous studies (e.g., Dutch-English: Tokowicz et al., 2002; Spanish-English: Prior et al., 2007) because of the conceptual and morphological mapping differences between the two languages. The bidirectional translation norms are the first translation equivalents database that examines Malay-English translation ambiguity index. Malay and English translations were gathered from two groups of proficient Malay-English speaking bilinguals to examine the number of possible correct translations for translations between the language pair. At word level, translation ambiguity was discussed alongside lexical and semantic characteristics, such as word class, within-language semantic variability, word length and frequency. At participant level, the relationship between bilinguals' language proficiency and translation accuracy was investigated.

Taken together, the translation norms served as a reference to aid stimuli selection for the translation tasks used to validate the newly developed LexMAL.

Chapter 3 describes the development and validation of LexMAL in two experiments. Because Chapter 2 revealed a need for measuring Malay L1 proficiency, Chapter 3 aimed at developing a valid and reliable Malay unspeeded yes/no vocabulary test that could measure the proficiency of both L1 and L2 speakers. The LexMAL prototype was constructed with 90 words and 90 nonwords that were carefully selected based on a set of predetermined criteria. In Experiment 1, the 180 items in the prototype were evaluated using point-biserial correlations and Item Response Theory analysis. The convergent validity of LexMAL was evaluated using Malay-English translations, cloze test, and self-rated proficiency. Based on the performance of Malay L1 and L2 speakers in Experiment 1 and 2, LexMAL's validity is discussed by comparing the LexMAL scores of Malay L1 and L2 speakers, and the correlations between LexMAL scores and accuracy of translation and cloze tests. In addition, the usefulness of objective and subjective language proficiency measures is discussed by comparing the correlations of LexMAL scores and self-rated proficiency with accuracy of translation and cloze tests.

Because LexMAL measures L1 and L2 vocabulary knowledge using a word recognition task (McLean et al., 2020), bilinguals' meaning and form recognition knowledge are expected to predict their yes/no vocabulary test scores better than meaning and form recall knowledge. However, unlike other form-meaning vocabulary tasks (e.g., form recognition tasks), there is no direct demonstration of form-meaning knowledge in the yes/no vocabulary tests like LexMAL. Therefore, it is unclear how test takers' form-meaning vocabulary knowledge can be inferred from their yes/no vocabulary test scores. Chapter 4 seeks direct evidence for how these written vocabulary knowledge (e.g., the vocabulary knowledge required for word recognition and recall) is used to answer LexMAL items. The contribution of bilinguals' form-meaning vocabulary knowledge to their item accuracy on LexMAL was examined using four vocabulary tests that were developed to assess different levels of form-meaning knowledge, namely meaning recognition, form recognition, meaning recall, and form recall. Language dominance and form-meaning knowledge level effects on LexMAL item accuracy were investigated using a generalised mixed-effects model. Furthermore, the contribution of form-meaning knowledge at each mastery level to LexMAL scores was investigated using a fixed-effects hierarchical regression analysis. The validity of LexMAL score as a vocabulary test and the role of form-meaning vocabulary knowledge in performing yes/no vocabulary test is further discussed using prediction from both models.

Chapter 5 summarises the findings of each chapter and discusses the contributions of the thesis to the literature. The theoretical contributions from assessing the translation performance of Malay-English bilinguals will be discussed to inform language-specific and language universal processing. Furthermore, the valuable insights from developing and validating a Malay proficiency test will be discussed to inform future test development for understudied languages. The utility of yes/no vocabulary test in measuring L1 and L2 lexical proficiency will be discussed in relation to a wide range of language proficiency measures employed in this thesis, including self-rated proficiency, translation production, and form-meaning vocabulary tests. In light of the theoretical framework reviewed in this introductory chapter, this thesis provides an account of test development for measuring both L1 and L2 proficiency. The process includes identifying the target language component of testing, evaluating the advantages and drawbacks of different test formats in evaluating the target language construct (Bialystok, 2001), selecting items systematically, and assessing the efficacy of these selection criteria (Hulstijn, 2015). The chapter will end with a discussion of recommendations for future research direction.

Chapter 2

Malay-English Bidirectional Translation Norms

Translation equivalents are widely used in bilingual research concerning word processing (e.g., Eddington & Tokowicz, 2013; Jouravlev & Jared, 2020) and second language vocabulary learning (e.g., Bracken et al., 2017; Degani et al., 2014). Although translation norms exist in several languages, there are yet no Malay-English translation norms. This chapter introduces Malay-English bidirectional translation norms, a new database developed as a part of this thesis to aid in the creation of translation tasks for LexMAL validation (see Chapter 3). The translation norms were gathered from highly proficient Malay-English bilinguals. Alongside the collection of Malay-English translation norms and ambiguity, the present study also investigated the impact of lexical characteristics on translation ambiguity. This chapter is based on the paper:

Lee, S. T., van Heuven, W. J. B., 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*. https://doi.org/10.3758/s13428-022-01977-3

2.1 Introduction

Studies investigating bilingual language processing often use translation equivalents, which are words from two different languages that share similar meaning (e.g., Basnight-Brown et al., 2020; Y. Lee et al., 2018). For example, Malay word "*beras*" and English word "rice grain" are Malay-English translation equivalents, in which both words refer to the seeds of a swamp grass that are cooked and consumed for food. The process of identifying appropriate translation equivalents requires researchers to be proficient in both languages, so that the meaning of the source word can be adequately represented in the translation. However, not all researchers are necessarily proficient in the languages of investigation (e.g., non-native Malay speakers conducting research in Malay). Furthermore, challenges in identifying translation equivalents are complicated by many words that do not have one-to-one corresponding translation from one language to another (Schwieter & Prior, 2020). For instance, the Malay word "*angka*" can be translated into "number", "digit", and "figure" in English, and thus have one-to-many mapping from Malay (source language) to English (target language). This one-to-many mapping from a source language to a target language is termed as translation ambiguity (Prior et al., 2007; Tokowicz et al., 2002).

2.1.1 Translation Ambiguity

Translation ambiguity could be driven by several reasons (Degani & Tokowicz, 2013; Prior et al., 2011; Schwieter & Prior, 2020). For example, translation ambiguity happens when meanings of a source word can be represented by different translations in the target language (e.g., Malay homonyms "*mangga*" can be translated into "mango", a type of fruit, and "lock", a tool that keeps door fastened, in English), or when a specific meaning of a source word (e.g., "*batu*" that refers to the solid substance found in the ground) can be translated into several possible English translations that share similar meanings (e.g., synonyms "rock" and "stone"). In addition, the conceptual and morphological differences (e.g., the use of affixations to signal meaning) between a language pair also contribute to the degree of translation ambiguity between two languages (Degani et al., 2016; Prior et al., 2007). For instance, the English word "thick" covers the meaning of "not thin" for both solid and liquid substances, however these concepts are distinctly represented by two words – "*tebal* (for solid)" and "*pekat* (for liquid)" in Malay.

2.1.2 Translation Norming Studies

There is no psycholinguistic database that could provide translation ambiguity index for every word in any given language pairs (Schwieter & Prior, 2020). Nevertheless, there were several translation norming studies conducted to estimate the prevalence of translation ambiguity for some language pairs. In these studies, bilinguals were asked to provide translations for words across the two languages they speak (e.g., Prior et al., 2007; Tokowicz et al., 2002; Wen & van Heuven, 2017a). Researchers then proceeded to identify the translation unambiguous (source words that have one-to-one mapping with its translation equivalents in the target language) and translation ambiguous words. Intriguingly, not all possible translations of the translation ambiguous words share the same status. In particular, the translation that is most frequently provided by bilinguals is identified as the dominant translation (e.g., the Spanish word "*permitir*" is a more dominant translation choice for the English word "answer" compared to the word "*dejar*"; Prior et al., 2007). The existence of dominant translations and translation norms enable researchers to further investigate factors affecting word translation and bilingual word processing

(Schwieter & Prior, 2020), such as how consistency of translation choice could be affected by translation ambiguity (e.g., Prior et al., 2011), and how the translation dominance of words (i.e., dominant and subordinate translations) affects bilingual language performance (e.g., in a translation recognition task: Bracken et al., 2017; Laxén & Lavaur, 2010).

Previous translation norming studies have demonstrated high translation ambiguity across several language pairs (Allen & Conklin, 2014; Prior et al., 2007; Tokowicz et al., 2002; Tseng et al., 2014; Wen & van Heuven, 2017a). The prevalence of translation ambiguity varies across different language pairs and translation directions (see Table 2.1 for summary). The observed differences in the prevalence of translation ambiguity could partially be attributed to methodological differences across studies (e.g., different sets of word stimuli, different number of participants) and unique language-specific linguistic characteristics (e.g., morphological complexity) (Schwieter & Prior, 2020). Furthermore, within each language pair, forward translation (L1-to-L2 translation, see Section 1.3.3.5 for a review) consistently resulted in lower translation ambiguity when compared to backward translation (L2-to-L1 translation; Allen & Conklin, 2014; Prior et al., 2007; Tokowicz et al., 2002)[1]. However, in view of (a) the current lack of variety in the language pairs being normed, and (b) all the available norms shared English as one of the languages, the extent of how language-specific linguistic characteristics contribute to translation ambiguity remains speculative.

[1] It is important to note that in all the reviewed translation norming studies, English is consistently being used as the target language in forward translation, and source language in backward translation.

Table 2. 1

Summary of translation ambiguity from past translation norming studies

Translation Norms	FT		BT	
	N	%	*N*	%
Dutch-English (Tokowicz et al., 2002)	562	25.3	562	30.4
Spanish-English (Prior et al., 2007)	762	48.2	670	58.5
Chinese-English (Tseng et al., 2014)	-	-	562	67.3
Chinese-English (Wen & van Heuven, 2017a)	-	-	1,429	71.2

Note. Tokowicz et al. (2002) and Tseng et al. (2014) normed on the same set of English words. The "first translation" method was used in all the translation norming studies above.

In the past, many bilingual studies assumed the words they used were translation unambiguous (Tokowicz et al., 2002). Such assumption can be problematic for interpreting research findings because studies have shown that bilinguals' performance on linguistic tasks can be affected by the degree of translation ambiguity (Eddington & Tokowicz, 2013; Jouravlev & Jared, 2020; Laxén & Lavaur, 2010). Specifically, bilinguals were found to recognise translation unambiguous word pairs faster than translation ambiguous word pairs, and the dominant translations were recognized faster than the non-dominant translations (see Schwieter & Prior, 2020, for a review). In such a scenario, translation norms are crucial for selecting translation equivalents for psycholinguistics studies investigating bilingual language processing.

Despite the growing number of studies investigating cross-linguistic word processing in Malay (e.g., Luniewska et al., 2019; M. J. Yap et al., 2010; N. T. Yap et al., 2017), there are no translation norms for Malay and English, which are commonly used as language pair in Malay cross-linguistic research. Therefore, the selection of Malay-English translation equivalents is subject to possible unforeseen extraneous variables and biases. Hence, the present Malay-English translation norming project aimed to create the first freely available large database of Malay-

English and English-Malay translation norms using the “first translation” method. This method is commonly used in translation norming studies (e.g., Allen & Conklin, 2014; Prior et al., 2007; Tokowicz et al., 2002; Tseng et al., 2014; Wen & van Heuven, 2017a), whereby participants are required to provide the first translation that comes to mind for each source word presented, resulting in a set of unique translations for each source word. The number of unique correct translations as determined by bilingual dictionaries is then used to calculate the translation ambiguity index for each source word (Schwieter & Prior, 2020). This translation norming project started with the forward translation phase that included 1,004 Malay words before these words were translated back from English to Malay in backward translation phase. Separate groups of proficient Malay-English bilinguals were recruited for each phase. The translations gathered were summarised into ambiguous and unambiguous translation equivalents, supplemented with word class, semantic variability (number of senses), word frequency and word length information. The availability of this information also allows further investigations into how lexical and semantic factors as well as individual differences might affect translation ambiguity and bilinguals’ translation choice.

2.1.3 Factors Affecting Translation Ambiguity

2.1.3.1 Word Class

Past studies suggest that verbs impose greater processing demands than nouns due to the complex relationship between semantics, syntax, and morphology of verbs (see Vigliocco et al., 2011 for a review). In general, nouns refer to discrete entities while verbs refer to actions or events. When comparing nouns and verbs within a language, meaning of verbs is often more context-dependent (Earles & Kersten, 2017; Gentner, 1981) and more polysemous (Miller & Fellbaum, 1991). Nouns across languages also have stronger conceptual overlap and are perceived to be more concrete than verbs in general (Bultena et al., 2013; Gentner, 1981; Laxén & Lavaur, 2010; Peti-Stantić et al., 2021; van Hell & de Groot, 1998). In some languages, members of a word class can be morphologically more complex than the others. For instance, whereas only English verbs can be inflected with different markers to indicate tenses and direction of actions, both Malay nouns and verbs can be inflected with several forms of affixes to form new words. These common irregularities of verbs could cause behavioural uncertainties and delay processing efficiency during language tasks (e.g., Maziyah Mohamed et al., 2023). Unlike past translation norming studies that mostly focused on nouns (e.g., Tokowicz et al., 2002; Wen & van Heuven, 2017a), the present translation norms also include words from other word classes (e.g., verbs and adjectives). The only translation norming study that compared translation ambiguity across different word classes (Spanish-English: Prior et al., 2007) revealed that verbs were significantly more translation ambiguous than nouns in both translation directions. In addition to nouns and verbs, the present study sets out to also compare the translation ambiguity of words from other grammatical classes, namely adjectives and class-ambiguous words.

2.1.3.2 Within-Language Semantic Variability

Previous translation norming studies also reported that within-language semantic variability, or words with multiple related senses within a language, are likely to be translation ambiguous (Allen & Conklin, 2014; Degani et al., 2016). In addition, the dominant meaning of the source words was more frequently translated in the translation, compared to its subordinate meanings (Degani et al., 2016). For instance, different Malay translation equivalents are possible for the English word "big" because it has two senses, with "*besar*" refers to the size of an object (i.e., large/not small), and "*penting*" refers to the importance of an event (i.e., important). Taking meaning dominance into account, "*besar"* is expected to be the dominant translation for the English word because it carries the dominant (more common) meaning of the word. In cases of speeded translation tasks, the Distributed Conceptual Feature Model (de Groot, 1992; van Hell & de Groot, 1998) suggested that semantic information of source words is shared by both source and target languages across nodes at the level of semantic representation. The more senses (meaning) a word carries, the more semantic nodes are available and may be activated in the semantic representation level. Thus, translations that share more semantic nodes with the source word would be more activated and subsequently speed up the translation process compared to other translations that share less semantic overlapping.

Employing senses information from official Malay dictionaries published by Dewan Bahasa and Pustaka Malaysia, the government body responsible for coordinating the use of the Malay language and literature in Malaysia, the present study investigated the effects of within-language semantic variability on translation ambiguity, as well as meaning dominance probability in the translations.

2.1.3.3 Word Length and Word Frequency

Previous translation norming studies have shown that word length and word frequency affect translation ambiguity. However, the effects were inconsistent across studies and dependent on which language pairs were involved (Prior et al., 2007; Tseng et al., 2014; Wen & van Heuven, 2017a). For example, Prior et al. (2007) found that low frequency words were more translation ambiguous than high frequency words in both Spanish-English and English-Spanish translation. In contrast, the opposite finding was observed for English-Chinese translations, more frequent English words were inclined to have more Chinese translations (Tseng et al., 2014; Wen & van Heuven, 2017a). Furthermore, Wen and van Heuven (2017a) also found that word frequency affected translation choice, where high frequency English words tended to be translated into high frequency Chinese translations.

Word frequency and word length effects are also predicted by bilingual word processing models that account for speeded translation accuracies and latencies (e.g., Multilink: Dijkstra et al., 2019). However, it is important to note that translation in speeded tasks is different from ("offline" or unspeeded) translation production studied in our and other translation norming studies. For instance, in the Multilink model (Dijkstra et al., 2019), word frequency affects the activation of word candidates in online translation production, where more frequent word candidates are activated faster than the less frequent ones. These unconscious and automatic activations draw on one’s implicit word knowledge (Durrant et al., 2022). The activation of word candidates is also expected to be stronger and more effortless if they share orthographic similarity with the source words (Dijkstra et al., 2019). If these findings also apply to offline translation tasks where bilinguals are asked to provide the first translation that come into their mind, translation candidates of high word frequency and with similar word length as the source words should be provided as the translation more readily. However, it is unclear yet whether the predictions of speeded responses in Multilink can be extended to offline tasks that require bilinguals to make conscious and controlled decisions from their explicit word knowledge. Moreover, the current development of Multilink assumes that semantic representations are mutually shared across languages. Put differently, the model does not yet account for partial semantic overlapping among translation equivalents (Dijkstra et al., 2019). Further research is required to address this issue because, in the lack of language-dependent semantic features, bias could be introduced into the prediction. Because the present study focused only on offline tasks and response time was not recorded, our findings will be discussed in light of Multilink with caution. The gathered translation norms, therefore, could serve as a foundation for future research to investigate the generalisability of Multilink's prediction to offline translation.

Negative word length effects, where shorter words tended to be more translation ambiguous, had been observed when English source words were translated into Spanish, but not when translating in the other direction (i.e., Spanish-English; Prior et al., 2007), nor when a different target language was involved (i.e., English-Chinese; Tseng et al., 2014). This finding is surprising because longer English words with seemingly lower word frequencies should be more translation ambiguous (Sigurd et al., 2004). Unfortunately, Prior and colleagues (2007) did not offer any explanation for this negative word length effect. Because Prior et al.'s (2007) study was the only study that showed the negative word length effect, it could be the result of the specific language pair and their unique cross-linguistic interactions. Future research is needed to fully understand this novel finding, given that the negative word length effect has not been replicated with other language pairs. If the word length effect can only be replicated with the same language pair but not the other language pairs, the unique cross-linguistic interactions from English to Spanish can be confirmed.

2.1.3.4 Individual Differences

In addition to the semantic and lexical effects on translation ambiguity, previous translation norming studies also reported individual differences in the translation word choice. Prior et al. (2007) revealed that more proficient L2 speakers were more consistent at producing the dominant translation (i.e., translation choice made by majority of the participants), although the effect was observed only in forward translation. Interestingly, L2 proficiency was shown to be correlated with translation accuracy and translation choice in the English-Chinese backward translation norms when L2 proficiency was estimated by an objective language proficiency measure, LexTALE (Lemhöfer & Broersma, 2012), but not when subjective self-rated proficiency was employed (Wen & van Heuven, 2017a). Taken together, previous studies revealed an influence of L2 proficiency on translation (effect sizes ranged from r = .39 to r = .51), whereby bilinguals with higher L2 proficiency are more likely to achieve greater agreement in the translation choice. Expectations of the language proficiency effects on translation performance can differ depending on the translation direction (Laufer & Aviad-Levitzky, 2017; Laufer & Goldstein, 2004; Schwieter & Prior, 2020). For instance, forward translation involves the process of form recall of L2 words. Bilinguals who are less proficient in their L2 are usually expected to have smaller L2 vocabulary size, thus might face greater L2 word retrieval difficulties when translating from their L1 to their L2, resulting in lower translation accuracy in forward translation. Conversely, meaning recall of L2 words is involved when bilinguals were translating words from their L2 to L1. In this case, less proficient bilinguals might not have complete semantic representation for the L2 source words, leading them to translate the only meaning that they know (which might not be the dominant meaning), and hence, showing lower agreement on the translation choice.

2.1.4 The Malay-English Translation Norms

The present Malay-English translation norms gathered correct translations in forward and backward translation directions. For each source word, the index of translation ambiguity (the number of distinct translations that matched with the meanings in dictionaries) and the dominant translations agreed by the majority were identified. In-line with previous translation norming studies (Prior et al., 2007; Wen & van Heuven, 2017a), bilinguals with higher L2 proficiency were expected to perform better in the translation tasks and more likely to provide a translation that matches the dominant translations provided by the majority than bilinguals with lower L2 proficiency. Furthermore, factors underlying translation ambiguity, translation choice and translation accuracy were examined. In line with most past research, we expected greater translation ambiguity for verbs, adjectives and word class ambiguous items when compared to nouns (Prior et al., 2007), and source words with higher number of senses to be more translation ambiguous, with a higher tendency for the dominant meaning of the source words to be provided as the dominant translation (Allen & Conklin, 2014; Degani et al., 2016). In addition, we explored the relationship between Malay lexical characteristics (word length and word frequency) and translation ambiguity. Translation equivalents were expected to resemble lexical characteristics of the source words, in which frequent words were expected to yield frequent dominant translations and longer words were expected to yield longer dominant translations (in line with Wen & van Heuven, 2017a).

2.2 Method

2.2.1 Participants

Sixty proficient Malay-English bilinguals were recruited. Half of the participants (11 males and 19 females) performed the forward translation from Malay to English, and the other half (9 males and 21 females) performed the backward translation from English to Malay. To investigate the translation ambiguity of forward and backward translation for the same set of words, the English source words for backward translation were selected from the translations produced by bilinguals in forward translation (following Prior et al., 2007, in investigating the translation ambiguity of the same set of words). Therefore, the recruitment was conducted in two phases, with participants for the forward translation phase recruited before the participants for the backward translation phase. All participants self-identified themselves as Malay-dominant speakers and were students studying at the University of Nottingham Malaysia. Participants were informed that the dominant language is operationalised as the most frequently used language in daily life and the language that participants find themselves to be most proficient in (Treffers-Daller, 2016). All participants met the English proficiency entry requirement of the university (IELTS Academic overall score 6.5 or equivalent). They received course credits or monetary compensation for their participation.

Participants completed a language background questionnaire adapted from the Language History Questionnaire 3 (P. Li et al., 2020) to report their language history, as well as their self-rated language proficiency in Malay and English on a scale from 1 (*very poor*) to 7 (*native-like*). All participants were early bilinguals who reported to have learnt Malay prior to English. In the forward translation phase, 28 participants acquired English by five years old, and the other two by seven years old; twenty-seven and three participants acquired English by the ages of five and seven, respectively, in the backward translation phase. Paired sample *t*-tests revealed that there was no significant difference between participants' self-rated Malay and English proficiency, *t*s ≤ 1.69, *p*s ≥ .10, suggesting that participants were highly fluent in both languages.

In addition to self-rated proficiency, participants' LexTALE (Lemhöfer & Broersma, 2012) scores confirmed that they were intermediate (*n* = 7 with 60% - 80% accuracy) - advanced (*n* = 53 with >80% accuracy) English users[2] (Lemhöfer & Broersma, 2012). Importantly, participants in the forward translation and backward translation phases were matched in terms of

[2] At the point of time when this research was conducted, there was no freely available objective Malay vocabulary measure (e.g., Malay version of LexTALE). Therefore, only English vocabulary knowledge was measured objectively.

their self-rated Malay and English proficiency as well as LexTALE score, *t*s ≤ 1.63, *p*s ≥ .11. A summary of the language background questionnaire and the LexTALE scores is presented in Table 2.2.

Table 2. 2

Summary of language background questionnaire and LexTALE data

	Forward Translation		Backward Translation	
	Mean	*SD*	*Mean*	*SD*
Age (years)	21.33	2.12	21.33	3.74
Age exposed to Malay (L1) (years)	0.43	1.02	0.10	0.54
Age exposed to English (L2) (years)	2.17	2.35	2.08	2.24
Self-rated L1 proficiency				
Reading	6.47	0.67	6.47	0.72
Writing	5.67	0.87	5.53	1.18
Listening	6.53	0.72	6.67	0.65
Speaking	6.23	1.12	6.27	1.03
Average	6.23	0.64	6.23	0.74
Self-rated L2 proficiency				
Reading	6.37	0.66	6.20	0.70
Writing	5.70	1.00	5.73	0.77
Listening	6.20	0.75	6.33	0.60
Speaking	5.73	0.89	5.93	0.73
Average	6.00	0.65	6.05	0.56
LexTALE score	87.21	8.57	90.42	6.26

Note. LexTALE (Lemhöfer & Broersma, 2012); Language background questionnaire measured self-rated proficiency on a 7-point scale (1 = *very poor*, 7 = *native-like*).

2.2.2 Stimuli

The present study used the Malay Lexicon Project (M. J. Yap et al., 2010) database as the main corpus for lexical information of Malay words, and SUBTLEX-US (Brysbaert et al., 2012; Brysbaert & New, 2009) for lexical information of English words. Word concreteness ratings for

English words[3] were taken from Brysbaert et al. (2014), ranging from 1 (*abstract*) to 5 (*concrete*). Zipf scale (van Heuven et al., 2014) was used as the word frequency measure instead of frequency per million words because it offers a more intuitive interpretation for users. Zipf values given to lexical items range from 1 (*very low frequency*) to 7 (*very high frequency*), with the boundary between low frequency and high frequency words lying between 3 and 4 (van Heuven et al., 2014). These categorization labels allow users to identify lexical items base on their frequency categories. Because the Malay Lexicon Project (M. J. Yap et al., 2010) provides only frequency count per million words, the Zipf value for each Malay word was calculated using the equation below (adapted from van Heuven et al., 2014).

$$\text{Zipf value} = \log_{10}\left(\frac{\text{Frequency count per million words} + 1.0}{\text{Corpus size in millions} + \text{number of word types in millions}}\right) + 3.0$$

$$= \log_{10}\left(\frac{\text{Frequency count per million words} + 1.0}{2.14 \text{ millions} + 0.009592 \text{ millions}}\right) + 3.0$$

The 1,004 Malay words involved in forward translation were selected from the 1,520 words used in M. J. Yap et al.’s (2010) lexical decision and speeded pronunciation experiments. This subset of words included 570 words that originated from 190 morphemic triplets. Each triplet contained a root word (e.g., “*hidup*/live”), its noun-affixed form (e.g., “*penghidupan*/life”) and verb-affixed form (e.g., “*menghidupkan*/give life”). Of these 570 words, 498 words were excluded to ensure every word appeared only once in the word list, either in root word form or affixed form. Root words were retained whenever possible, and affixed words with the highest word frequency were kept in cases where root words were absent. In the example given above,

[3] In view of concreteness rating database was not available for Malay words, only concreteness ratings for English words were made available for the present study.

root word “*hidup*” was kept and its affixed forms - “*penghidupan*” and “*menghidupkan*” were removed. Lastly, these words were checked against a Malay-English dictionary (*Kamus Melayu-Inggeris Dewan,* Jasmani, 2012) to identify and exclude words that have sole culture-specific (e.g., “*joget*/a type of Malay dance”) or religious meaning (e.g., “*iblis*/devil”) because they do not have a direct translation in English.

The final word set (1,004 Malay words) had a mean word frequency (in Zipf value) of 3.94 (*SD* = 0.73) and a mean word length of 6.80 (*SD* = 2.59) in the Malay Lexicon Project (M. J. Yap et al., 2010). Word class information obtained from *Kamus Perdana* (S. H. Cheng & Lai, 2019) revealed that the word set comprised of 374 nouns, 228 verbs, 116 adjectives, 278 word-class ambiguous items (e.g., “*aksi*” can be a noun or an adjective), four adverbs, one classifier, one pronoun, one numeral, and one interjection. The Malay words were randomly split into 10 blocks of 100 words (except for one block that had 104 words). Words in the blocks were matched in Zipf value and word length. One sample *t*-tests conducted against the average Zipf value (*M* = 3.94, *SD* = 0.73) and the average word length (*M* = 6.80, *SD* = 2.59) revealed no significant differences with individual word block’s Zipf value and word length, *t*s ≤ 1.12, *p*s ≥ .27.

After English translations for the 1,004 Malay words were gathered in the forward translation phase, all correct dominant single-word English translations were used as stimuli for the backward translation phase. For Malay words that received no correct translation, or correct dominant translations that have more than one word in forward translation, the expected single-word English translations from the reference Malay-English dictionary (*Kamus Melayu-Inggeris Dewan*, Jasmani, 2012) were used. Malay words with no single-word English translations according to the Malay-English translation norms and the reference dictionary were excluded (*n*

= 12). Furthermore, the English translations that appeared more than once in the forward translation norms were presented only once in backward translation phase (e.g., "level" was the dominant English translation for Malay words "*darjat*", "*paras*", and "*peres*", and it was presented only once in backward translation phase). The final backward translation stimuli set consisted of 845 English words.

Overall, the English word stimuli had a mean word frequency (Zipf value) of 4.26 (*SD* = 0.91), mean word length of 6.20 (*SD* = 2.27), and mean concreteness ratings of 3.25 (*SD* = 0.97). To match with the word class classification of the Malay words in the forward translation task, we utilized the *all part-of-speech*[4] information for English word class (Brysbaert et al., 2012). There were 123 nouns, 94 verbs, 46 adjectives, 576 word class ambiguous items, four adverbs, one determiner, and one interjection (see Table 2.3 for a comparison of word class across translation directions). The English words were randomized into 9 blocks of 100 words (except for the final block that had 45 words). One sample *t*-tests confirmed that words in the blocks were matched in Zipf value, word length and concreteness, *t*s ≤ 1.82, *p*s > .07.

[4] There are two types of word class classification available for the English words (Brysbaert et al., 2012), namely *All part-of-speech* (All_PoS) and *dominant part-of-speech* (dom_PoS). For instance, the word "float" was observed as a noun (105 times) and a verb (276 times) in Brysbaert et al. (2012). Consequently, "float" is classified as word-class ambiguous under All_PoS because more than one word class is associated with the word. In addition, verb is listed as the dom_PoS of "float" in view of its higher frequency of occurrence when compared to the noun form.

Table 2. 3

Distribution of word class across translation directions

Word class	Forward Translation (Malay Words)	Backward Translation (English Words)
	n	*n*
Noun	374	123
Verb	228	94
Adjective	116	46
Word class ambiguous	278	576
Adverb	4	4
Interjection	1	1
Pronoun	1	-
Numeral	1	-
Determiner	-	1
Classifier	1	-

2.2.3 Procedure

In the forward translation phase, participants translated 4 blocks of words every day and completed the translation task in 3 days within a week. The presentation of word blocks within a day and words within each block was randomized. The word stimuli were presented in lowercase, one word at a time, as black characters on a silver background using PsychoPy (Peirce et al., 2019). Participants were required to enter the first translation that came to their mind. They could skip items by pressing the ENTER key if they could not provide a translation. After finishing each block, participants were prompted to take a short break. On the third day of translation, the LexTALE test (Lemhöfer & Broersma, 2012) and language background questionnaire were administered on Qualtrics (https://www.qualtrics.com), after participants had completed the final two blocks of words. The same procedure was adopted for backward translation phase, except that the backward translation participants translated 3 blocks of words a day and completed the 845 translations in 3 days within a week. The experiment was approved by the Ethics Committee in the School of Psychology at the University of Nottingham Malaysia. Written consent was acquired from participants before data collection started.

2.2.4 Scoring

Translation accuracy of participants was determined by comparing their translations against the expected translations provided by the Malay-English and English-Malay dictionaries. For the expected Malay-English translations, *Kamus Melayu-Inggeris Dewan* (Jasmani, 2012) was used as the primary reference source, and *Kamus Perdana* (S. H. Cheng & Lai, 2019) was used as the secondary reference. For English-Malay translation, *Kamus Dwibahasa* (Ibrahim, 2002) was chosen as the primary reference while Oxford English-English-Malay Dictionary (Oxford University Press & Oxford Fajar, 2018) was used as the secondary reference. The primary reference dictionaries were selected because they are widely used by Malay language users as the official dictionary in Malaysia. They were published by the Institute of Language and Literature, the official government body that monitors Malay language development and usage in the country.

Grammatical affixations that did not transform the word class of a word, such as third person singular ‘-s’ and plural ‘-s’ in English, were collated to its root word and accepted as correct responses if they matched the expected translations. Spelling errors were corrected and accepted on the condition that the errors did not result in another real word in the target language. Two proficient Malay-English coders further examined the translations that did not match with the expected dictionary translations. Synonyms of the expected dictionary translations and colloquial meanings provided were further examined and coded as correct responses only upon agreement achieved from both coders. Some judgment criteria used to accept exceptional translations included: (a) the translations shared similar meaning as the expected translations provided by the dictionaries and both could be used interchangeably (e.g., "*siap*" was accepted as a synonym for "*habis*" and "*selesai*" because both carry the meaning of "finish"), and (b) translations matched with the word choice used colloquially in daily conversations (e.g., "*orang*" as a translation for "human"). Responses that described the meaning of the source words instead of being the direct translation were rejected (e.g., "hairless" for "*botak*/bald").

2.3 Results

This section first describes the Malay-English bidirectional translation norms and the translation ambiguity index gathered (see Sections 2.3.1 and 2.3.2). After that, the roles of language proficiency, source word frequency and word length in influencing translation accuracy were explored using Spearman's rho and Wilcoxon signed ranked tests in Section 2.3.3. Section 2.3.4 further investigated source words that received at least one correct translation investigated to determine the roles of word class, within-language semantic variability, word frequency and word length in translation ambiguity. Finally, Section 2.3.5 examined the effects of meaning dominance, word frequency and word length on bilinguals' translation word choice. Semantic and lexical information of all the words gathered in the translation norms can be found in the Open Science Framework (OSF) repository[5].

[5] https://osf.io/cnkjq/?view_only=54b5521c763241faa18a5b70963f2550

2.3.1 Malay-English Forward Translation Norms

2.3.1.1 Translation Accuracy

The forward translation phase resulted in a total of 27,130 English translations (90.1%) and 2,990 omitted responses (9.9%). A total of 18,378 translations (67.7%) were correct responses. Of the 1,004 Malay words, 64.2% (645 words) were correctly translated by at least 50% of the participants, 31.4% (315 words) received correct translations from at least one participant, and 4.0% (44 words) of the stimuli received no correct translation.

2.3.1.2 Translation Ambiguity

Translation ambiguity was determined by the number of possible translations provided for each source word. When a source word yielded only one unique correct translation, it was considered as translation unambiguous, and a source word was considered translation ambiguous when it resulted in more than one correct translation. In the forward translation norms, the number of possible translations provided for the Malay words ranged from zero to eight. Of the 1,004 Malay words, 63.3% of them were translation ambiguous words (see Table 2.4). Across the ambiguity range, 45.4% of the translation ambiguous words had two unique correct translations, 28.0% had three unique correct translations, and 26.6% had four or more unique correct translations (see Figure 2.1 for the distribution of words according to their translation ambiguity).

2.3.1.3 Dominant Translations

For translation unambiguous words, the unique translation equivalents are the dominant translations. The dominant translations for the translation ambiguous words were identified by selecting the correct translations that were most frequently provided by the participants. In cases where the translation ambiguous words had more than one dominant translation (48 words, 4.78%), the translation that matched with the dominant meaning from the primary reference dictionary was selected. The results revealed that the dominant English translations of the forward translation norms covered a wide range of word lengths (*M* = 6.20, *SD* = 2.65, minimum = 2, maximum = 23), word frequencies (Zipf value) (*M* = 4.37, *SD* = 0.90, minimum = 1.59, maximum = 7.62) and concreteness ratings (*M* = 3.25, *SD* = 0.95, minimum = 1.19, maximum = 5).

2.3.2 English-Malay Backward Translation Norms

2.3.2.1 Translation Accuracy

The backward translation phase resulted in 23,813 Malay translations (93.9%) and 1,537 omitted responses (6.1%). Of the Malay translations, 20,454 responses were correct translations (85.9%). Overall, 88.4% (747 words) of the 845 English words received correct translations from at least 50% of the participants, 10.9% (92 words) were translated correctly by at least one participant, and 6 words (0.7%) received no correct translation from the participants.

2.3.2.2 Translation Ambiguity

The number of possible translations in the backward translation norms ranged from zero to eleven, with 78.0% of the 845 English words being translation ambiguous (see Table 2.4). Of these translation ambiguous words, 39.5% of the translation ambiguous words had two unique correct translations, 24.6% had three unique correct translations, and 36.0% had at least four unique correct translations (see Figure 2.1 for the distribution of translation ambiguous words).

The translation ambiguity of backward translation was compared against the forward translation norms using the same set of 845 source words used in both translation directions. In forward translation phase, 34.2% (289 words) of these words were translation unambiguous while 62.4% (527 words) were translation ambiguous. The numerical percentages suggest that English-Malay backward translation resulted in more translation ambiguity compared to forward translation (see Figure 2.1).

2.3.2.3 Dominant Translations

In the backward translation norms, 29 translation ambiguous words (3.43%) had more than one dominant translation. Overall, the dominant Malay translations had a mean word length of 6.80 (*SD* = 2.49, minimum = 3, maximum = 19), and mean word frequency (Zipf value) of 4.17 (*SD* = 0.75, minimum = 2.83, maximum = 6.63).

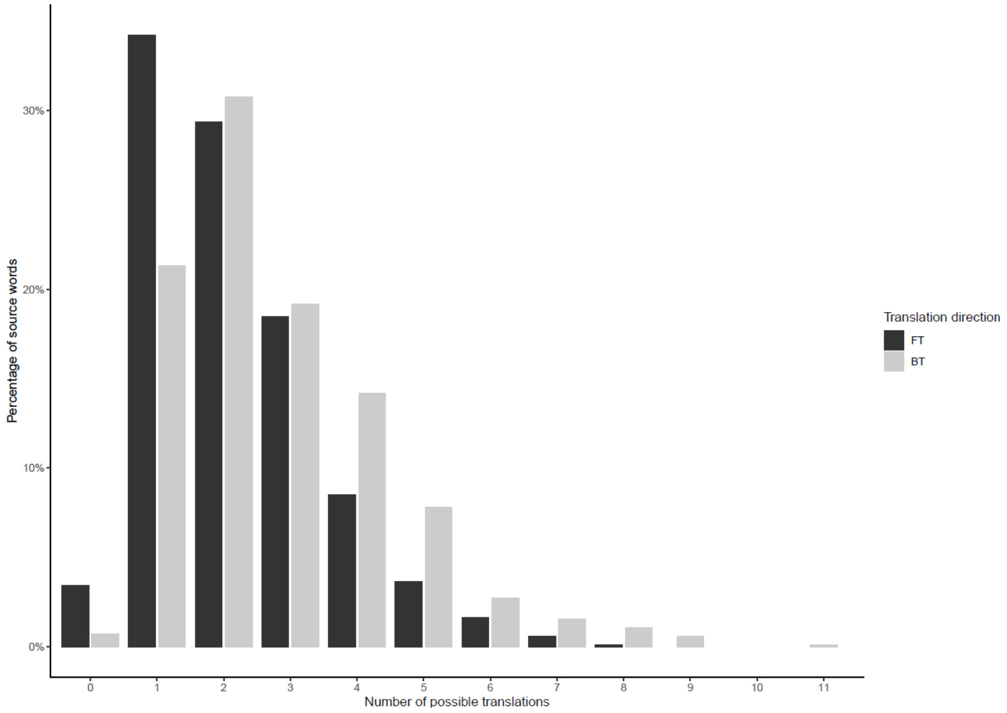
Table 2. 4

Proportion of Malay and English words according to their translation ambiguity for the Malay-English and English-Malay translation norms

Type of translation pair	Number of source words	Proportion (%)
Malay-English forward translation norms (*N* = 1,004)		
Translation unambiguous	325	32.4
Translation ambiguous	635	63.3
No correct translation	44	4.4
English-Malay backward translation norms (*N* = 845)		
Translation unambiguous	180	21.3
Translation ambiguous	659	78.0
No correct translation	6	0.7

Figure 2. 1

Distributions of the 845 Malay and English words according to their number of possible translations for the Malay-English forward translation and English-Malay backward translation norms



2.3.3 Translation Accuracy

The set of analyses reported in this section assessed factors that affect bilinguals' translation accuracy. The role of language proficiency was investigated at participant level, followed by word length and word frequency analyses at both participant and item levels. Dominant translation scores were determined based on the percentage of correct dominant translations each participant provided (participant level) or gathered for each source word (item level), and translation accuracy score was calculated by computing the percentage of correct translations made in total independent of whether the translation was dominant or non-dominant.

Before examining the factors affecting translation accuracy, Shapiro-Wilk tests of normality were conducted and revealed non-normal distribution of the translation scores (*p*s < .01). Therefore, non-parametric tests were conducted for this set of analyses.

2.3.3.1 Language Proficiency

At participant level, the influence of language proficiency on translation performance of proficient Malay-English bilinguals was investigated.

In forward translation, Spearman's rho test revealed a statistically significant moderate, positive correlation between self-rated L1 Malay proficiency and participants' dominant translation scores, as well as translation accuracy scores (see Table 2.5). Participants who perceived themselves as having higher Malay proficiency provided more dominant translations and more correct translations. However, L2 proficiency measures (i.e., LexTALE and self-rated English proficiency) did not correlate with these translation scores, *p*s > .09. Interestingly, none of the language proficiency measures in the backward translation group correlated with participants' translation scores, *p*s > .50.

Table 2. 5

Spearman's rho (r_s) for language proficiency and translation accuracy

Variable	Dominant translation score	Translation accuracy score
FT (*N* = 30)		
LexTALE score	.09	.13
Self-rated L1 proficiency	.49**	.38*
Self-rated L2 proficiency	.25	.31
BT (*N* = 30)		
LexTALE score	-.08	.08
Self-rated L1 proficiency	.13	.10
Self-rated L2 proficiency	-.04	.09

Note. *Correlation is significant at the 0.05 level (two-tailed).

**Correlation is significant at the 0.01 level (two-tailed).

2.3.3.2 Language Entropy

A language entropy analysis (Gullifer & Titone, 2020) was conducted to examine diversity of language use among the bilinguals and its impact on bilinguals' dominant translation scores and translation accuracy scores. Language entropy scores for the participants across four common social settings (i.e., communication with family, friends, course mates and communication in other social contexts) were computed using the language entropy R package (Gullifer & Titone, 2018). With a language entropy value range from 0 (i.e., no language diversity and only one language was being used across the four social contexts) to 1 (i.e., high language diversity and balanced use of the two languages across different contexts), the forward translation participants revealed a similar mean language entropy value of .78 (*SD* = .29) as the backward translation participants who revealed a mean language entropy value of .79 (*SD* = .28), *t*(58) = -0.14, *p* = .89. The language entropy analysis indicates that the bilingual participants commonly used two languages (i.e., Malay and English) in the four measured social contexts.

Importantly, Spearman’s rho indicated that participants’ individual language entropy scores do not correlate with their dominant translation scores and translation accuracy scores, *p*s ≥ .24.

2.3.3.3 Word Length and Word Frequency

The effects of word length and word frequency on translation accuracy were examined at participant and item levels. At participant level, Wilcoxon signed rank tests were conducted to compare the translation accuracy of high and low frequency words as well as long and short words in both translation directions. Source words with a Zipf value of 4 and above were considered as high frequency words, and source words with Zipf value below 4 were considered as low frequency words. At the same time, the source words from each direction were split into two groups around the mean word length (mean word length for forward translation = 7.00; backward translation = 6.21). Table 2.6 summarised the proportion of source words in each lexical group.

Wilcoxon signed rank tests revealed that the translation accuracy of the high frequency words was significantly higher than that of low frequency words in both translation directions, *p*s < .001 (see Figure 2.2). Also, the translation accuracy for shorter words was significantly higher than that of longer words, *p*s ≤ .007 (see Table 2.7). Overall, participants demonstrated higher translation accuracy and were more likely to provide dominant translation for high frequency and short words, in contrast to low frequency and long words.

Table 2. 6

Proportion of source words according to lexical characteristics

Variable	Forward translation		Backward translation	
	N	*Proportion (%)*	*N*	*Proportion (%)*
Word frequency				
High	428	42.63	525	62.13
Low	576	57.37	320	37.87
Word length				
Long	300	29.88	209	24.73
Short	704	70.12	636	75.27

Table 2. 7

Wilcoxon signed rank tests to compare translation accuracy by lexical characteristics

Variable	Dominant translation score			Translation accuracy score		
	T	*z*	Effect size (*r*)	*T*	*z*	Effect size (*r*)
FT (*N* = 30)						
Word frequency	465	-4.78***	.87	465	-4.78***	.87
Word length	102	-2.68**	.49	87	-2.99**	.55
BT (*N* = 30)						
High vs low frequency words	465	-4.78***	.87	465	-4.78***	.87
Long vs short words	23	-4.31***	.79	465	-4.78***	.87

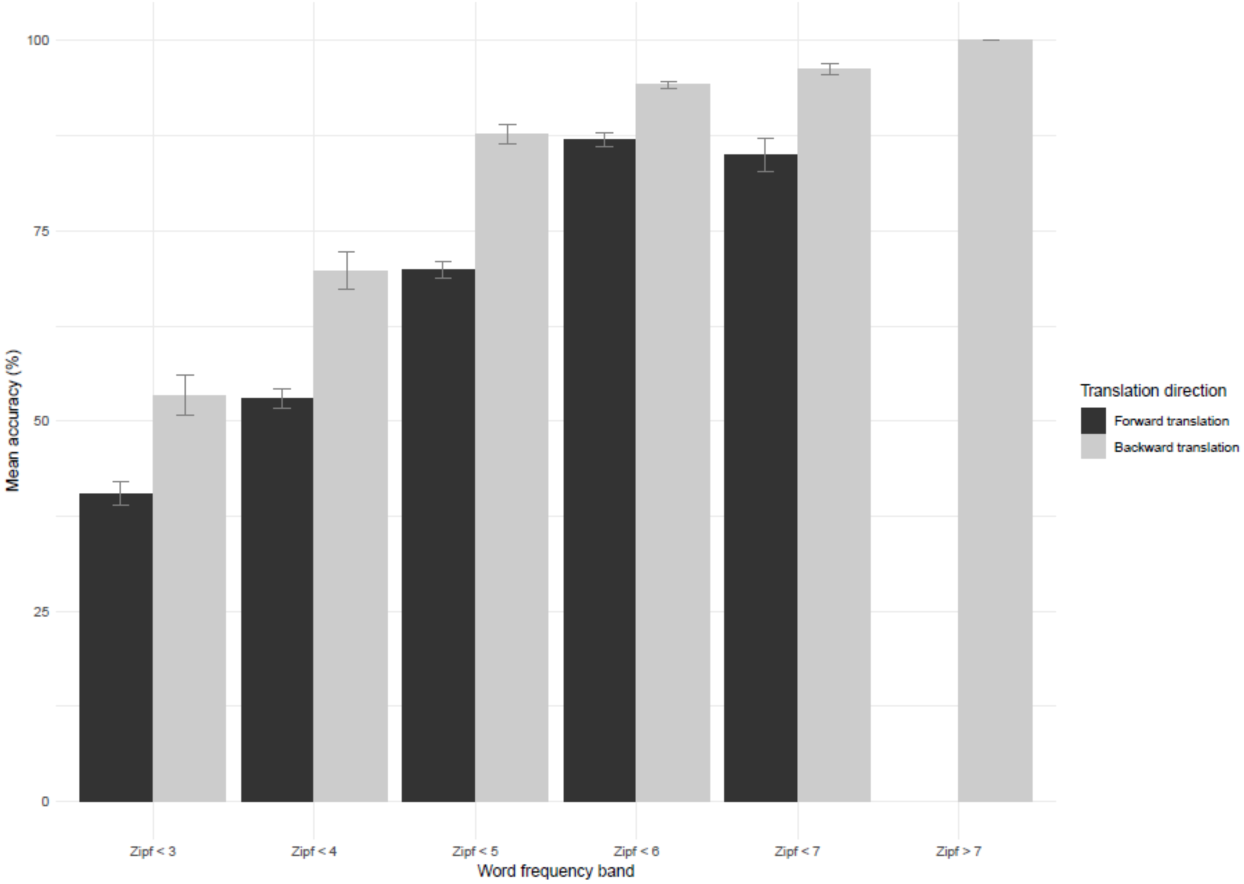
Note. Effect size in *r* was converted from *z* score (Clark-Carter, 2019).

** Difference was significant at the <.01 level (two-tailed).

*** Difference was significant at <.001 level (two-tailed).

Figure 2. 2

Translation accuracy according to word frequency bands



Spearman's rho was conducted to assess the subsequent relationships between source words' lexical characteristics and translation performance. In both translation directions, source words' frequency positively correlated with dominant translation and translation accuracy scores, while source words' length negatively correlated with both dominant translation and translation accuracy scores, *p*s < .001 (see Table 2.8). To establish the direction of effects, linear regression models were fitted to estimate the proportion of variance in translation accuracy explained by source words' frequency and length in each translation direction. The models explained a significant 15% and 33% of the variance in participants' forward (*F*(2, 1001) = 91.27, *p* < .001) and backward translation accuracy (*F*(2, 842) = 209.43, *p* < .001). In forward translation, translation accuracy was positively predicted by source words' frequency, *B* = 17.99, *SE* = 1.40, *t*(1001) = 12.89, *p* < .001, and negatively predicted by word length, *B* = -0.90, *SE* = 0.39, *t*(1001) = -2.31, *p* = 0.021. In backward translation, however, only source words' frequency significantly predicted translation accuracy, *B* = 13.84, *SE* = 0.81, *t*(842) = 16.99, *p* < .001.

Table 2. 8

Spearman's rho (r_s) for source words' lexical characteristics and translation accuracy

Variable	Dominant translation score	Translation accuracy score
1,004 Malay source words		
Word frequency	.32***	.39***
Word length	-.15***	-.17***
845 English source words		
Word frequency	.42***	.62***
Word length	-.30***	-.43***

Note. ***Correlation is significant at the 0.001 level (two-tailed).

2.3.3.4 Language Proficiency on Words with Matched Word Frequency and Length

Comparing the word frequency of source words in both translation directions revealed that the mean word frequency (in Zipf values; van Heuven et al., 2014) of Malay source words in forward translation (*M* = 3.98, *SD* = 0.73) was significantly lower than that of English source words in backward translation (*M* = 4.27, *SD* = 0.91), *t*(1612.88) = -7.25, *p* < .001. A closer look at the proportion of high and low frequency words also revealed that more than half of the forward translation source words (57.37%) were low frequency words with Zipf value less than 4, whereas only 37.87% of the backward translation source words were of low frequency. Therefore, the word frequency difference between forward and backward translations could be a confounding factor in the difference in translation accuracy observed between the two tasks and for the L1 proficiency effect observed in forward translation.

Therefore, additional analyses were conducted to investigate whether the L1 proficiency effect and the lexical effects on translation accuracy remained the same when word frequency and word length in both tasks were matched. The LexOPS R package (Taylor et al., 2020) was used to generate a subset of 709 words with word frequency (mean for forward translation: 4.13, 43.72% low frequency words; mean for backward translation: 4.20, 40.76% low frequency words) and length (mean for forward translation: 6.51; mean for backward translation: 6.36) matched (*p*s ≥ .12) between the two translation directions.

At participant level, with the subset of carefully matched 709 words, we found again significant correlation between self-rated L1 Malay proficiency with dominant translation scores, $r_s = .47$, $p = .009$, and the translation accuracy scores, $r_s = .40$, $p = .03$. No significant correlation was found between all the proficiency measures with translation accuracy in backward translation, *p*s ≥ .43.

At item level, the correlation results of the subset replicated that of the full set. The findings again revealed significant positive correlations between source words' frequency with dominant translation scores (forward translation: $r_s = .34$, backward translation: $r_s = .38$; *p*s < .001) and translation accuracy scores (forward translation: $r_s = .43$, backward translation: $r_s = .56$; *p*s < .001). In addition, significant negative correlations were found between source words' length with dominant translation scores (forward translation: $r_s = -.15$, backward translation: $r_s = -.24$; *p*s < .001) and translation accuracy scores (forward translation: $r_s = -.19$, backward translation: $r_s = -.36$; *p*s < .001). Regression analyses indicated that only word frequency predicted translation accuracy ($B = 18.82$, $SE = 1.63$, $t(706) = 11.54$, $p < .001$) in the forward translation model ($R^2 = 0.17$, $F(2, 706) = 74.77$, $p < .001$), while both word frequency ($B = 14.65$, $SE = 1.03$, $t(706) = 14.17$, $p < .001$) and length ($B = -0.68$, $SE = 0.34$, $t(706) = -1.99$, $p = .05$) predicted translation accuracy in the backward translation model ($R^2 = 0.27$, $F(2, 706) = 133.38$, $p < .001$).

In sum, the effects of language proficiency on forward translation performance, as well as word frequency on translation performance in both directions remained significant when the word frequency and length were matched between both tasks. Word length effects, on the other hand, became smaller or even negligible when the lexical characteristics were matched in both translation directions.

2.3.4 Translation Ambiguity

2.3.4.1 Word Class

To investigate if translation ambiguity was affected by word class, source words from each translation direction were grouped by four distinct word classes: nouns, verbs, adjectives, and word class ambiguous items. Source words that belong to other word classes (i.e., adverb, classifier, determiner, interjection, numeral and pronoun) were excluded from this analysis because the sample size for each of these word classes was too small to generate meaningful comparisons (see Table 2.9 for word class distribution). A Kruskal-Wallis ANOVA[6] indicated that there were significant differences across translation ambiguity of nouns, verbs, adjectives, and word class ambiguous items, H (corrected for ties) = 27.85, $df = 3$, $N = 952$, $p < .001$,

[6] Because translation ambiguity was not normally distributed (as indicated by Shapiro-Wilk test of normality, $p < .001$), non-parametric tests were conducted.

Cohen's *f* = .17. Separate Mann-Whitney *U* post-hoc tests revealed that translation ambiguity for nouns was significantly lower than that of verbs, adjectives, and word class ambiguous items, *p*s < .005. There was no significant difference across the translation ambiguity of verbs, adjectives and word class ambiguous items, *p*s ≥ .18. Table 2.10 presents the post-hoc tests' results.

Table 2. 9

Translation ambiguity index according to word class in forward translation

Word class	*N*	TA (%)
Nouns	374	54.8
Verbs	228	70.2
Adjectives	116	68.1
Word-class ambiguous	278	67.3

Note. 996 words retrieved from the 1,004 Malay source words. TA = translation ambiguity.

Table 2. 10

Post-hoc Mann-Whitney U tests to compare translation ambiguity across word class in forward translation

Word class	*U*	*z* (corrected for ties)	*p*	Effect size (*r*)
Nouns vs verbs	29661	-4.80***	.000	.20
Nouns vs adjectives	16574	-2.84**	.004	.13
Nouns vs word-class ambiguous	39372	-3.74***	.000	.15
Verbs vs adjectives	11570	-.93	.351	
Verbs vs word-class ambiguous	27227	-1.33	.184	
Adjectives vs word-class ambiguous	15066	-.08	.936	

Note. Effect size in *r* was converted from *z* score (Clark-Carter, 2019).

** Difference was significant at the <.01 level (two-tailed).

*** Difference was significant at <.001 level (two-tailed).

Similar word class analyses were conducted on the 845 English words in the backward translation (see Table 2.11 for word class distribution). Kruskal-Wallis ANOVA confirmed that there were significant differences across translation ambiguity of nouns, verbs, adjectives, and word class ambiguous items, *H* (corrected for ties) = 36.89, *df* = 3, *N* = 833, *p* < .001, Cohen's *f* = .22. Mann-Whitney *U* post-hoc tests revealed that verbs were significantly more translation ambiguous than nouns, adjectives, and word class ambiguous items, *p*s ≤ .02. At the same time, adjectives and word class ambiguous items were significantly more translation ambiguous than nouns, *p*s ≤ .05. There was no significant difference between translation ambiguity of adjectives and word class ambiguous items, *p* = .28 (see Table 2.12 for summary).

Table 2. 11

Translation ambiguity index according to word class in English

Word class	*N*	TA (%)
Nouns	123	63.4
Verbs	94	90.4
Adjectives	46	71.7
Word-class ambiguous	576	79.5

Note. 839 words retrieved from the 845 English source words. TA = translation ambiguity.

Table 2. 12

Post-hoc Mann-Whitney U tests to compare translation ambiguity across word class in backward translation

Word class	*U*	*z* (corrected for ties)	*p*	Effect size (*r*)
Nouns vs verbs	3104	-5.80***	.000	.40
Nouns vs adjectives	2255	-1.97*	.048	.29
Nouns vs word-class ambiguous	24836	-5.04***	.000	.19
Verbs vs adjectives	1620	-2.38*	.017	.20
Verbs vs word-class ambiguous	22617	-2.40*	.017	.09
Adjectives vs word-class ambiguous	11936	-1.09	.275	

Note. Effect size in *r* was converted from *z* score (Clark-Carter, 2019).

* Difference was significant at the 0.05 level (two-tailed).

*** Difference was significant at < .001 level (two-tailed).

2.3.4.2 Within-language Semantic Variability

The relationship between within-language semantic variability and translation ambiguity was further investigated in this section. Semantic variability was defined by the number of senses (meaning) a word has according to the primary reference dictionary. All possible meanings associated with a particular word form were summed up, including meanings of homonyms (words that share the same form but carry distinct meanings, e.g., "*guna*" was considered to have three senses, namely the two related senses "use" and "role", as well as the (third) unrelated sense "spell"). Nineteen Malay words from forward translation and eight English words from backward translation were excluded from the analysis because their number of senses were not provided by the primary reference dictionary. Non-parametric Spearman's rho tests indicated statistically significant positive correlations between the number of senses of words and number of possible translations in FT, $r_s = .23$, $p < .001$, two-tailed, $N = 951$, and BT, $r_s = .25$, $p < .001$, two-tailed, $N = 833$. Words with higher semantic variability tend to have higher number of possible translations.

2.3.4.3 Word Length and Word Frequency

Spearman's rho conducted indicated weak, yet statistically significant positive correlation between Malay word length and the number of translations provided, $r_s = .08$, $p < .05$, two-tailed, $N = 960$. Similarly, Malay word frequency also correlated weakly and positively with the number of translations provided, $r_s = .09$, $p < .01$, two-tailed, $N = 960$. Malay words with longer strings and of higher frequency were more likely to yield more translations. The same correlation analyses conducted for the English words in backward translation however, only word length showed a trend towards a positive correlation with the number of translations provided, $r_s = .06$, $p = .06$, two-tailed, $N = 839$.

2.3.5 Translation Word Choice

The next analyses investigated the effects of meaning dominance, word frequency and word length on translation word choice. Only translation pairs for which at least 50% of the participants provided the dominant translations were included in the following analyses to ensure that the translations under investigation truly represent the translation choice of the majority of the participants.

2.3.5.1 Meaning Dominance

This section focuses on the roles of semantic and lexical characteristics in bilinguals' translation word choice. The probability of meaning dominance effect, defined by the likelihood for the dominant meaning of a source word (as indicated by the primary reference dictionary) to also be a dominant translation, was first examined. For instance, the effect was demonstrated when most of the participants translated the English word “direction” into its dominant meaning “*arah*”, rather than its sub-dominant meaning “*arahan*”.

Of the 502 Malay translation ambiguous words in forward translation, 405 Malay source words had their dominant meaning translated by majority of the participants, and 97 words had their sub-dominant meaning translated by the majority. A chi-square test for goodness of fit was conducted to assess if the dominant meaning of source words were more frequently translated than the sub-dominant meaning. The chi-square test revealed that the frequency of the dominant meaning being translated into dominant translation was significantly higher than that of the sub-dominant meaning, χ^2 (1, N = 502) = 188.97, $p < .001$ (Cohen’s w = 0.61).

For the 576 English translation ambiguous words in backward translation, 341 had their dominant meaning translated by majority of the participants, and 235 had their sub-dominant meaning translated by the majority. The dominant meanings of English words, when compared to subdominant meanings, were also more frequently translated into the dominant Malay translations, $\chi 2$ (1, N = 576) = 19.51, $p < .001$ (Cohen’s w = 0.18).

2.3.5.2 Word Length and Word Frequency

The present study also examined the relationship between word length of the source words and their dominant translations. In forward translation, Spearman’s rho test revealed a relationship between the word length of Malay source words and English translations, r_s = .31, p < .001, two-tailed, N = 502, indicating that longer Malay words were translated into longer English words. Similarly, there was also a statistically significant correlation between Malay and translated English word frequency, r_s = .41, p < .001, two-tailed, N = 502, indicating that more frequent Malay words were translated into more frequent English words.

In backward translation, a significant correlation was also found between the word length of English source words and Malay translations, r_s = .49, p < .001, two-tailed, N = 576. The positive correlation indicates that longer English words were translated into longer Malay words. Before proceeding to the word frequency correlational analysis, an additional 39 English-Malay translation pairs were excluded because the word frequency information was not available for the Malay translations. Spearman's rho indicated a moderate yet statistically significant positive correlation between the word frequency of English source words and Malay translations, r_s = .46, p < .001, two-tailed, N = 537. In other words, more frequent English words were translated into more frequent Malay translations.

2.4 General Discussion

The present study aimed at creating the first freely available Malay and English translation norms with proficient Malay-English bilinguals. As a result, a database of Malay-English and English-Malay translation norms for 1,004 Malay words and 845 English words is formed. The norms predominantly consist of nouns, verbs, adjectives and class ambiguous words that span across a range of semantic variability, word frequencies and word length. Section 2.4.1 discusses translation ambiguity between Malay and English in relation to other language pairs studied in previous translation norming studies. Factors affecting translation choice and translation ambiguity are discussed in sections 2.4.2 and 2.4.3, respectively.

2.4.1 Translation Ambiguity

The Malay-English forward translation norms revealed a high proportion of translation ambiguous Malay words (63.3%). This proportion is higher compared to other translation norms that also involved English as the target translation language (e.g., Dutch-English: 25.3%, Tokowicz et al., 2002; Spanish-English: 48.2%, Prior et al., 2007). The exceptionally low translation ambiguity reported in the Dutch-English norms are likely an underestimation because the stimuli were chosen and assumed to be translation unambiguous by previous research (Schwieter & Prior, 2020). In contrast, the Malay source words used in this study were not selected based on being translation unambiguous. Similarly, the English-Malay backward translation norms also revealed high translation ambiguity between the two languages (78.0%), which was higher compared to other backward translation norms (e.g., English-Dutch: 30.4%, Tokowicz et al., 2002; English-Spanish: 58.5%, Prior et al., 2007), even when compared to the English-Chinese translation norms in which the two languages are differently scripted (67.3% in Tseng et al., 2014; 71.2% in Wen & van Heuven, 2017a).

We attributed the high translation ambiguity observed in the present study to the conceptual mapping differences between Malay and English. Malay as an Austronesian language and English as an Indo-European language come from two different language families. In comparison to language pairs that belong to the same language family group (e.g., Dutch and English which are both varieties of West-Germanic languages of the Indo-European language family), Malay and English are likely to have relatively more distinct concepts for words (Schwieter & Prior, 2020; Tseng et al., 2014). Translation ambiguity could emerge when a source language has a wide conceptual space for words (e.g., “thick” for both solid and liquid), whereas the target language provides finer distinctions for the concepts (e.g., “*tebal*” for solid and "*pekat*" for liquid). In such case, a single concept carried by a source word can result in two different translations in the target language.

On top of that, we also found translation ambiguity of English-Malay backward translation norms to be higher than the Malay-English forward translation norms. This finding is consistent with past translation norming studies (Prior et al., 2007; Tokowicz et al., 2002), in which translation from English as a source language to another target language (e.g., English-Dutch) always resulted in higher translation ambiguity compared to translation in the other direction (e.g., Dutch-English). Because the higher translation ambiguity has been observed with English as the source language, it is likely that the language-specific properties of English, such as greater within-language semantic variability (Degani et al., 2016), contributed to the higher number of possible translations in the target languages. In addition, the morphological mapping differences between English and Malay could have added to the variability in translation too, with English being morphologically less complex than Malay. As an example, the English word "need" can be translated into different forms of Malay word "*perlu*", including the root word "*perlu*", verb-affixed form "*memerlukan*", and noun-affixed form "*keperluan*".

The higher translation ambiguity and translation accuracy observed in backward translation compared to forward translation could also be due to the L2-L1 translation direction because bilinguals were translating from their less dominant language to their more dominant language in backward translation (Schwieter & Prior, 2020). These bilinguals were likely to be more proficient in Malay than English because they were self-identified as Malay L1 and dominant speakers, even though their self-rated language proficiency for the two languages did not differ significantly. If we assume a larger vocabulary size in the bilinguals' L1 (Rahman et al., 2018), more translation choices would be available for translation equivalents in L1, compared to when translation was conducted in the other direction. However, as far as we are aware of, all existing backward translation norms use English as the source language, hence it is not possible to pinpoint the higher translation ambiguity in backward translation to language-specific properties (e.g., polysemous English) or language-universal factor (e.g., better vocabulary knowledge in the target language). Thus, future backward translation studies could consider to (a) employ a source language other than English to provide additional evidence regarding the role of language-specific characteristics of the source language in translation ambiguity (Schwieter & Prior, 2020), and (b) recruit bilinguals who speak English as their L1 or dominant language to perform the same translation task. If the source language of a backward translation task has a narrower conceptual space (Schwieter & Prior, 2020) than the target language, and yet still results in higher translation ambiguity than the forward translation task, the L2-L1 effect explanation on translation ambiguity (language-universal factor) would be supported. If dominant or L1 English speakers performing in an English-Malay translation task (L1-L2 translation) show higher translation ambiguity than the Malay-English translation task, it would suggest that the translation ambiguity observed in the present study is likely to be induced by language-specific characteristics of the English language.

With respect to lexical factors affecting translation ambiguity, the present study replicated the findings from Prior et al. (2007) by showing that verbs were more translation ambiguous than nouns in both translation directions. In addition, adjectives and word-class ambiguous items were at least as translation ambiguous as verbs. However, it is important to note that the word class effects observed might be affected by the different proportions of words in each word class across the forward and backward translation phases. For instance, the higher proportion of class-ambiguous items in backward translation compared to forward translation might have enabled the study to reveal a lower translation ambiguity of class-ambiguous item when compared to verbs. Therefore, future research is needed to investigate the interplay between word class, translation direction, and translation ambiguity. Nevertheless, because verbs, adjectives and word-class ambiguous items were significantly more translation ambiguous than nouns, it is likely that the higher translation ambiguity found in the present study than in other translation norming studies involving mostly nouns (e.g., Allen & Conklin, 2014; Prior et al., 2007; Tokowicz et al., 2002) could be partly attributed to the additional word classes used. For instance, when the translation ambiguity of words from different word classes were taken into account, English words were more translation ambiguous with Malay (78.0% in the present study) than Chinese (67.3% in Tseng et al., 2014; 71.2% in Wen & van Heuven, 2017a). However, when only the translation ambiguity of nouns was considered, the translation ambiguity index of English-Malay translation became less ambiguous (63.4%) than the English-Chinese translation.

The present findings also replicated the positive relationship between within-language semantic variability and translation ambiguity in both translation directions. In the past, English words with more senses (high semantic variability) tend to produce a greater number of possible translations in Dutch, German, Spanish, and Hebrew (Degani et al., 2016). Although the present study investigated a different language pair in two translation directions, similar effects were found. Allen and Conklin (2014) also reported in their Japanese-English translation norming study a similar effect of semantic variability in both translation directions. This finding from offline translation appears to be in-line with the prediction of the Distributed Conceptual Feature Model (de Groot, 1992; van Hell & de Groot, 1998), in which the semantic information of source words is activated in the translation process, and the degree of ambiguity in translation depends on the semantic overlapping between the source and target languages (Laxén & Lavaur, 2010). Because within-language semantic variability appears to affect bilinguals' translation choice, bilingual research investigating cross-language processing should take semantic variability into consideration when selecting word stimuli. This information can be easily accessed via bilingual dictionaries.

In addition to the impact of the number of senses on translations, longer and frequent Malay words resulted in more translations in English. This is likely due to the conceptual and morphological differences (Degani et al., 2016; Prior et al., 2007; Schwieter & Prior, 2020) between Malay and English. For instance, the highly frequent Malay pronoun "*dia*" (Zipf value = 6.09) can be translated into gender-specific pronouns in English, namely "he" and "she", resulting in two unique correct translations. Moreover, the rich morphological system in Malay (see Section 1.4.1 for an introduction) might also cause longer Malay words to be more ambiguous for translation (e.g., "*perubatan*" can either be translated as "medical" or "medicine" in English).

Surprisingly, these lexical effects appear to be specific to the status of source and target language identity within the language pair. When the source and target language is swapped in the backward translation task, only English word length showed a trend towards a positive correlation with translation ambiguity, whereas word frequency effect of English source words was not observed. This finding corroborates with previous research findings, whereby word frequency and word length effects on translation ambiguity are found to be inconsistent and sensitive to the source and target language identity. For instance, for Spanish and English, word length effects on translation ambiguity became negligible when the source-target language was changed (i.e., from English-Spanish to Spanish-English) (Prior et al., 2007). Furthermore, the direction of word frequency effects could change when the source language remained the same and only the target language was substituted (e.g., negative correlations for English-Spanish translations but positive correlations for English-Chinese translations; Prior et al., 2007; Tseng et al., 2014; Wen & van Heuven, 2017a). In sum, it appears that the relationship between two languages could differ according to language-specific properties of the language pair in question. Because different English word sets were employed across these studies, it is difficult to pinpoint which factor contributed to the discrepancy.

The interpretation of existing evidence for lexical effects is further complicated by the translation direction involved. For instance, it remains unclear whether translation direction may have contributed to the unique lexical effects observed in each source-target language pair because most of the existing translation norming studies only investigated one translation direction for each source-target language pair (e.g., backward translation from English-Chinese: Tseng et al., 2014; Wen & van Heuven, 2017a; see Prior et al., 2007 for an exception). Although Prior and colleagues revealed that lexical effects for the same source-target language pair (e.g., Spanish-English) were consistent regardless of translation direction, further replication is needed to assess the generalisability of this finding. Therefore, future studies should consider using the same set of source words for meaningful cross-linguistic and cross-study comparisons, as well as investigating forward and backward translation with the same source words.

2.4.2 Translation Choice

The present study also replicated the meaning dominance effect whereby dominant meaning of source words provided in the primary dictionary is more likely to become the dominant translation (Degani et al., 2016). Although the dictionary we used provides a brief statement that the meanings of the vocabulary items are arranged according to the commonality of usage, to our knowledge, there is no empirical evidence yet that supports the dominance of the meanings first listed in it. The present findings provide the first preliminary evidence as such. The effect suggests consideration of the semantic overlapping between source words and translations is common during translation (Laxén & Lavaur, 2010). Although previous studies only investigated meaning dominance effects in backward translation, the findings from our study provide empirical evidence that meaning dominance effects occur in both translation directions.

Besides the consideration of meanings, further correlational analyses also revealed that in both translation directions longer source words were translated into longer words, and more frequent source words were translated into words with higher frequencies. These findings are in-line with previous translation norming studies that employed different language pairs (Japanese-English: Allen & Conklin, 2014; English-Chinese: Wen & van Heuven, 2017a), indicating that lexical characteristics of the source words have an influence on translation choice for any language pair and translation direction. These word frequency and word length effects observed in off-line or unspeeded translation appears to be consistent with the predictions of the online word processing model. For instance, Multilink (Dijkstra et al., 2019) predicted that activations for translations with higher word frequency or greater orthographic similarity to the source words would be stronger. As a result, when participants were asked to select the first translation that came to mind, translations with higher word frequency or similar word lengths to the source words were more likely to be selected. However, it is important to note that Multilink is designed to account for online word processing. Further research is needed to investigate the extent to which the predictions of speeded responses can be extended to offline translations.

2.4.3 Translation Accuracy and Language Proficiency

Only in forward translation that participants who rated themselves with higher Malay (L1) proficiency were more likely to provide correct and dominant translations. Surprisingly, this correlation was not found in backward translation. One possible explanation is that the overall word frequency of the Malay and English source words differed between the two translation directions. For the 845 source words shared by both translation directions, the mean word frequency (in Zipf values) of Malay source words in forward translation (*M* = 3.98, *SD* = 0.73) was significantly lower than that of English source words in backward translation (*M* = 4.27, *SD* = 0.91), *t*(1612.88) = -7.25, *p* < .001. A closer look at the proportion of high and low frequency words involved also revealed that more than half of the forward translation source words (57.37%) were low frequency words with Zipf value less than 4, while only 37.87% of the backward translation source words were of low frequency (see Table 2.6). Because low frequency words are expected to tap into the higher language cognition of L1 speakers (Hulstijn, 2015), this high number of low frequency words in forward translation could be a potential confound of the L1 proficiency effect observed, whereby high proficiency and vocabulary knowledge in L1 Malay became an important factor for participants to perform well in forward translation.

To investigate whether the L1 proficiency effect on translation accuracy remained when word frequency and word length in both tasks were matched, an additional analysis was conducted using a subset of 709 words that were carefully matched. These results revealed again a significant effect of L1 language proficiency on forward translation performance. This finding suggests that in addition to form recall knowledge in L2 (Laufer & Goldstein, 2004; Schmitt, 2010), meaning recall knowledge of the bilinguals in their L1 is also associated with their translation accuracy in forward translation.

To the best of our knowledge, no past translation norming study has investigated and revealed the impact of L1 proficiency on translation performance, probably because bilinguals' L1 proficiency was always assumed to be homogeneous as a group. In accordance with the Basic and Higher Language Cognition theory (Hulstijn, 2015, 2019), the present study provides preliminary evidence to point out that even though most bilingual studies assumed "native-speaker" proficiency (Izura et al., 2014), there could still be potential variation in L1 proficiency within a rather homogeneous group, and it could potentially influence L1 speakers' language performance (Diependaele et al., 2013). This might be especially true in a diverse multilingual society such as Malaysia. Future studies should consider extending the investigation of bilingual word processing to also include proficiency measures for L1, to account for possible language proficiency effects.

Surprisingly, in contrast to previous research, there was no correlation between L2 proficiency (indicated by objective LexTALE scores and subjective self-ratings) and translation word choice in forward and backward translation. Prior et al. (2007) found that Spanish L1 group with higher L2 proficiency were more likely to produce dominant translations, but only in forward translation. The impact of L2 proficiency was also found in the English-Chinese BT study (Wen & van Heuven, 2017a). However, it is important to note that these studies utilized different sets of stimuli and proficiency measures, which complicates direct comparison of findings across studies. Again, future studies should consider using objective L2 proficiency measure (e.g., LexTALE) and similar sets of source words for meaningful cross-study comparisons.

We suspect our bilinguals' high L2 competence to be one of the reasons why we did not find the relationship between L2 proficiency and translation accuracy. Most past translation norming studies recruited unbalanced bilinguals (e.g., Prior et al., 2007; Wen & van Heuven, 2017a), who reported to have learnt L2 in school and only later immersed in L2 environment during tertiary education. The present study however involved highly proficient bilinguals who have learned the L2 before attending school (< 7-year-old). Most of them had rated themselves to be equally proficient in Malay and English too, despite reporting Malay as their dominant language (cf. Duyck & Brysbaert, 2004). According to the theory of Basic and Higher Language cognition (see Section 1.1.2; Hulstijn, 2015, 2019), the highly proficient bilinguals in our studies are likely to have acquired basic English lexical knowledge that is commonly shared by English L1 speakers, and would show variation only in the higher language cognition. Because the English source words in the backward translation task were produced by highly proficient bilinguals in the forward translation task, they are probably well-known to most of our Malaysian proficient English speakers (as suggested by their higher word frequency when compared to the source words in the forward translation task). As a result, these relatively higher frequency words are not difficult or discriminative enough to distinguish the speakers' English proficiency (see Figure 2.2 for accuracy across frequency bands). In other words, to identify the variation in language proficiency of highly proficient speakers, the stimuli should contain a good blend of low and high frequency words (e.g., Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Wen et al. 2023).

Lastly, the present study also demonstrates that source words with higher word frequency and shorter word length were more likely to be translated correctly in both translation directions. These words seem to be easier items for the translation tasks. Correspondingly, Wen and van

Heuven (2017a) also found that their Mandarin-English bilinguals were more reliable in providing the dominant translations for high frequency English words. In Malay, longer words are likely to be words with affixations, which may or may not share the same word class with the root words (e.g., the root word "*hidup*/live" and one of its affixed form "*menghidupkan*/give life" are verbs; while another affixed form "*penghidupan*/life" is a noun). The uncertainties in word class of these longer Malay words with affixations could result in a higher chance of making translation mistakes, because participants have to first accurately identify the right meaning and word class form of the affixed words, before performing the translation. Taken together, our study provides evidence that word frequency and word length influence translation accuracy and hence can be used to estimate translation stimuli difficulty level for highly proficient Malay-English bilinguals. However, it is important to note that the present study collected translation equivalents only from highly proficient Malay L1 and English L2 speakers. Whether the findings can be generalised to other types of bilinguals with varying English L2 proficiency, or even Malay L2 speakers, remains to be tested.

2.5 Conclusion

The present study created the Malay-English and English-Malay translation norms through forward and backward translation tasks. The present translation norms are the first norms collected from highly proficient bilinguals. Our data analyses showed high prevalence of translation ambiguity between the Malay and English language and replicated some lexical characteristics and semantic variability effects on translation ambiguity. Although attempts to explain the inconsistency in these effects met with challenges due to the inconsistency in word stimuli used in past translation norming studies, we suggest standardising future norming items to help setting apart the language-specific and language-universal factors towards translation ambiguity.

The present translation norms provide the first database for researchers conducting language research with Malay-English bilinguals. Together with the lexical and semantic information of the source and target words, these norms provide a comprehensive reference to aid stimulus selection for future experimental studies (e.g., Jouravlev & Jared, 2020) and computer simulations (e.g., Dijkstra et al., 2019). In accordance with the Basic and Higher Language Cognition theory (Hulstijn, 2015, 2019), the present study demonstrated that there was some variation in L1 proficiency among the highly proficient bilinguals, which had an impact on their language performance when knowledge of low frequency words was tested. Building on the findings from this chapter, the next chapter developed a vocabulary test that can measure L1 and L2 proficiency.

Chapter 3

Lexical Test for Malay Speakers (LexMAL)

Objective language proficiency measures have been found to provide better and more consistent estimates of bilinguals’ language processing than self-rated proficiency (e.g., Tomoschuk et al., 2019; Wen & van Heuven, 2017a). However, objectively measuring language proficiency is often not possible because of a lack of quick and freely available language proficiency tests (Park et al., 2022). This chapter reports the process of developing and validating a Lexical Test for Malay Speakers (LexMAL), which estimates language proficiency for Malay L1 and L2 speakers. The LexMAL prototype was developed and validated in two experiments. Both experiments provided evidence to support the validity of LexMAL; LexMAL scores distinguished language proficiency of L1 and L2 speakers, and significantly correlated with translation accuracy and cloze test scores. Additionally, LexMAL scores outperformed self-rated proficiency in predicting translation and cloze test accuracy. This chapter incorporates material from the following paper:

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*. https://doi.org/10.3758/s13428-023-02202-5

3.1 Introduction

Language proficiency of bilinguals affects representations and processing of the languages they speak (see Jiang, 2015 for a review; see also Chapter 1.1). For instance, Chapter 2 revealed that the L1 proficiency of Malay-English bilinguals affects their translation accuracy, despite their proficiency is assumed to be homogenous as a group. In addition, bilinguals' performance in cross-linguistic tasks has also been found to be affected by L2 proficiency in previous studies (e.g., Sarrett et al., 2022; Wiener & Tokowicz, 2021). Therefore, when bilingual language processing is examined, previous experimental studies often measure language proficiency in L1 (Brysbaert et al., 2016; Diependaele et al., 2013; Hulstijn, 2015; see also Chapter 2) and L2 speakers (Diependaele et al., 2013; Wen & van Heuven, 2017a; H. Zhang et al., 2020). Objective language measures such as vocabulary size tests have been shown to provide reliable and accurate estimation of individual differences of language proficiency among bilinguals (e.g., Diependaele et al., 2013; Lemhöfer & Broersma, 2012; Tomoschuk et al., 2019; H. Zhang et al., 2020). However, systematic reviews (Park et al., 2022; Surrain & Luk, 2019) showed that objective language proficiency measures are not consistently used whenever language proficiency is measured, with less than 50% of bilingual research from the last decade using an objective language proficiency measure to assess participants' language proficiency.

One of the reasons why researchers rarely used objective language proficiency measures was that such tests are not freely available for many of the studied languages (Park et al., 2022). Furthermore, the use of some well-known standardised language proficiency tests might involve costs (e.g., International English Language Testing System, IELTS) or they take a long time to administer (e.g., 40 minutes for the Vocabulary Size Test, Nation & Beglar, 2007). The availability of objective language proficiency measures is especially rare for understudied languages. For instance, there is currently not freely available quick and standardised Malay proficiency test, although there are 377 million Malay speakers in the world.

In their effort to advocate for a standardised language measure across psycholinguistic studies, Lemhöfer and Broersma (2012) presented a yes/no unspeeded vocabulary test to measure the English proficiency of advanced learners of English and named it the Lexical Test for Advanced Learners of English (LexTALE; see Chapter 1.3.3.3 for an introduction). The LexTALE stimuli were selected from Meara's (1996) unpublished "10K" vocabulary size test via a pilot study. After collecting word/nonword decisions on all 240 items piloted, the difficulty level and discrimination power of each item were computed based on the percentage of accuracy and item-whole correlation, respectively. Four difficulty levels were formed separately for words and nonwords, and the items with the top 25% highest discrimination power from each difficulty level were selected for the LexTALE. A higher number of words than nonwords (2:1 ratio) are included in the test to ensure that the perceived proportion of words and nonwords is roughly equal, as some of the low frequency words may be subjectively interpreted as nonwords by test takers. After choosing the 60 LexTALE stimuli (40 words and 20 nonwords), an experiment (Lemhöfer and Broersma, 2012) that investigated the validity of the test was conducted. Participants for both the pilot and validation studies were recruited from the same population (i.e., Dutch-English bilinguals). Furthermore, a group of Korean-English speakers was also recruited for the validation study to investigate the utility of LexTALE for speakers of different L1s.

Four different measures, including L1-L2 translation (forward translation), L2-L1 translation (backward translation), the Quick Placement Test, and self-rated English proficiency, were used to assess the validity of LexTALE as a test of vocabulary. The LexTALE scores (i.e., mean percentage correct) correlated strongly with the other measures for the L1 Dutch group (*r*s ≥ .63) and weak-moderately for the L1 Korean group (*r*s ≥ .29). Building on these findings, the LexTALE has been commonly used in bilingual research as a measure of English vocabulary size for L2 speakers with advanced proficiencies (e.g., Diependaele et al., 2013; Wen & van Heuven, 2017a).

The widespread use of LexTALE has resulted in the consistency of language proficiency measurement in psycholinguistic research. To develop quick and valid language proficiency measures for non-English languages, its test format and construction have been extended to other languages (see Section 1.3.3.3 for the list of lextale-inspired tests). The lextale-inspired tests sought to measure L1 and L2 proficiency on the same scale (e.g., Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014), in contrast to LexTALE (and its parallel versions in Dutch and German, Lemhöfer & Broersma, 2012). Instead of being translated directly from the LexTALE, the word items for the lextale-inspired tests were carefully chosen based on word frequency in the target language to maximise their utility in differentiating between L1 and L2 speakers. The test prototypes typically consisted of a larger set of items and were tested with L1 and L2 speakers of the target population. Item assessment was conducted to select items that spread across the difficulty levels and had the highest discrimination power. To select reliable test items, point-biserial correlations between participants' accuracy on the test item and overall test scores was examined. Following that, item response theory (IRT) analysis was carried out to determine the discrimination power and difficulty level of each test item. With the ratio of words to nonwords kept at 2:1, each lextale-inspired test has its own item size (e.g., 60 words and 30 nonwords in LextPT, Zhou & Li, 2022). A follow-up validation study was conducted to evaluate the validity of the final set of items. Most studies (e.g., Brysbaert, 2013; Izura et al., 2014; Zhou & Li, 2022) only used self-reported measures (e.g., self-rated proficiency) to validate the tests. To the best of our knowledge, only one lextale-inspired test (i.e., Wen et al., 2023) was validated with external criterion measures, including translation tasks, cloze test, and self-rated proficiency. Therefore, LexTALE and its Dutch and German extensions are distinguished from the lextale-inspired tests that are developed in other languages because of the differences in difficulty level, item size, and validation procedures. Consequently, the scores obtained from these tests are not comparable across different languages, even though they share a common test format.

In addition to the differences in validation procedures, there are also some variations in the scoring systems utilised by the tests. The original LexTALE study (Lemhöfer & Broersma, 2012) assessed three different scoring systems, namely the mean percentage correct, the ΔM (Meara, 1992), and I_{SDT} (Huibregtse et al., 2002). The best scoring system for LexTALE was shown to be the mean percentage correct, which is calculated by averaging the percentages of correctly identified words and nonwords. The mean percentage correct, however, typically only ranges between 50% (i.e., chance level) and 100% (Brysbaert, 2013). For example, test takers who answered “yes” (or “no”) continuously throughout the test would receive a score of 50%, which is counterintuitive when it comes to test score interpretation. Therefore, Brysbaert proposed the Ghent score, which corrects the number of correct word identifications with the number of incorrect nonword identifications. With the Ghent score, a test taker who responds randomly to the test will obtain a Ghent score close to zero. Only test takers who know the words and correctly identify the nonwords would receive a high Ghent score. However, as pointed out by Wen et al. (2023), the Ghent score range depends on the number of words and nonwords included in the test, which differs across the lextale-inspired tests. To enable more transparent comparison in research that used more than one lextale-inspired test, Wen and colleagues proposed the use of normalised Ghent score (see equation shown below). It sums up the number of correctly identified words and penalises the score based on guessing by the participant ("yes" responses for nonwords, i.e., false alarms). The normalised Ghent score ranges from -100% to 100%, with a negative score indicating a higher false alarm rate than correct word identification.

$$\text{Normalised Ghent score} = \left(N_{yes\ to\ words} - 2N_{yes\ to\ nonwords}\right) \times \frac{100}{N_{words}}$$

To address the need of a reliable and valid quick Malay proficiency measure, we developed a lexical test for estimating language proficiency in Malay (LexMAL). Furthermore, we validated LexMAL with two external criterion measures: translation tasks and a cloze test. For the scoring of LexMAL, we used the normalized Ghent score (Wen et al., 2023). A receiver operator characteristic curve analysis was conducted as part of the evaluation of the validity of the test. Following previous studies (e.g., Wen et al., 2023), two experiments were conducted to construct and validate LexMAL. Experiment 1 (preparatory study) tested the LexMAL prototype to select the best items for the final LexMAL. The prototype was tested with two distinct groups of Malay speakers, namely Malay L1 (*N* = 60) and L2 (*N* = 60) speakers, to examine its ability to discriminate the two groups of Malay speakers based on their vocabulary size estimates.

As far as we are aware, there is no standardised Malay vocabulary test that can be used as the criterion comparison. Therefore, we included Malay-English bidirectional translations and cloze tasks as external criterion to validate LexMAL. In addition, Mandarin-Malay translation tasks were presented to Malay L2 speakers[7] to assess their Malay vocabulary knowledge in relation to their L1 (i.e., Mandarin Chinese, henceforth Mandarin). Bidirectional translation production tasks and the cloze test assessed different aspects of word knowledge. Translation production requires form recall knowledge, or the recall of word form in another language, whereas the cloze test assesses collocation knowledge, or how words should be used together. These tasks have been used by previous studies as criterion measures for vocabulary knowledge, and they have been found to consistently correlate with receptive vocabulary size (Lemhöfer & Broersma, 2012; Nakata et al., 2020; Wen et al., 2023; X. Zhang et al., 2020).

The final version of LexMAL was constructed based on item assessment conducted in Experiment 1. Following previous lextale-inspired tests that sought to measure L1 and L2 proficiency on the same scale (e.g., Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014), the final LexMAL, in contrast to LexTALE (Lemhöfer & Broersma, 2012), consists of larger stimuli set for higher test reliability and wider difficulty range (see Section 1.3.3.3). It consists of 60 words and 30 nonwords that cover a wide range of difficulty levels and at the same time demonstrated the greatest discriminatory power. In Experiment 1, sensitivity of the LexMAL prototype was examined by comparing LexMAL scores between the Malay L1 and L2 speakers, whereas its convergent validity was assessed by examining the correlations between LexMAL scores and participants' performance in the translation and cloze tasks. The validity evidence of the final LexMAL was evaluated in Experiment 2 (validation study). We expected Malay L1 speakers to score higher than L2 speakers in LexMAL, reflecting

[7] The Malaysian Chinese ethnic group makes up 24.6% of the Malaysian population and is the largest Malay L2 speaking ethnic group in Malaysia. They usually speak Mandarin as their L1 with some exceptions who speak other languages (e.g., English) as their L1.

the larger Malay vocabulary size expected in the L1 speakers. In addition, LexMAL was expected to show good internal reliability and good convergent validity and outperform self-ratings in predicting speakers' translation and cloze test scores.

3.2 Experiment 1: Preparatory Study

3.2.1 Method

3.2.1.1 Participants

While power calculations for test construction may differ from those for experimental designs, we were unaware of the method for calculating the required sample size for test construction. Therefore, we calculated the power specifically for the language group effect, which aligns with our primary hypothesis. The a priori power analysis conducted using G*Power (Faul et al., 2009) indicated that at least 51 participants were required for each language group to obtain .80 power to detect a medium effect size of .50 at the standard .05 alpha error probability. A medium effect size was used in the power analysis despite the fact that the effect sizes for L1-L2 proficiency differences were typically large, Cohen's $d \geq 2.91$ (e.g., Brysbaert, 2013; I. L. Chan & Chang, 2018). This is because Malaysian bilinguals are highly proficient in their L2, and therefore a smaller difference was expected between the Malay proficiency of L1 and L2 speakers. Following the smaller effect size (Cohen's $d = 0.7$) revealed in Ferré and Brysbaert's (2017) study that compared highly proficient L1 and L2 speakers, a medium effect size was used in the present study. The present study recruited a slightly larger sample than recommended to account for unforeseen issues in online studies such as incomplete surveys or dropouts. Sixty Malay L1 speakers (13 males and 47 females) and 60 proficient Malay L2 speakers (all spoke Mandarin as L1; 13 males and 47 females) were involved in this study. All Malay L1 speakers identified Malay as their L1 and dominant language (except for one who identified English as their L1, exposed to Malay at the age of 9 and continued to use Malay as their dominant language). All Malay L2 speakers but four (who reported to have been exposed to Mandarin and Malay simultaneously during childhood) 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, $t(118) = 10.60$, $p < .001$. The L1 speakers, on average, acquired Malay at a significantly younger age compared to the L2 speakers, $t(117.08) = 15.28$, $p < .001$. Table 3.1 summarises the speakers' language background collected using the same questionnaire as the translation norming studies described in Chapter 2.

All participants recruited were current students or graduates of tertiary education and had a minimum "Pass (C)" qualification for the *Bahasa Melayu* (Malay) and *Bahasa Inggeris* (English) subjects in the national high school examination (commonly known as the *Sijil Pelajaran Malaysia*, SPM). Participants received monetary compensation for their participation.

Table 3. 1

Summary of participants' language background

Variable	Malay L1			Malay L2		
	Range	*Mean*	*SD*	Range	*Mean*	*SD*
Age (years)	20–38	23.92	3.21	20–44	25.82	4.75
Age of acquisition (years)						
Malay	0–9	0.50	1.56	0–10	5.05	1.70
English	0–7	4.60	2.25	0–8	4.28	2.09
Mandarin				0–7	0.57	1.63
Self-rated proficiency						
Malay	2.50–7.00	6.39	0.86	2.75–6.50	4.80	0.77
English	2.50–7.00	5.15	0.81	3.50–7.00	5.03	0.82
Mandarin				4.50–7.00	6.14	0.73

Note. Language background questionnaire measured self-rated proficiency on a 7-point scale (1 = *very poor,* 7 = *native-like*).

3.2.1.2 Stimuli

The present experiment involved five tasks to assess different Malay language skills and to collect self-rated language proficiency and language background information. Details of the stimuli used in each of these five tasks are described in the following subsections. Instructions were presented in English throughout the study, except for the instructions used in the LexMAL prototype, which were presented in Malay. The tasks and the items within each task were presented in the same order to all participants.

Task 1: LexMAL Prototype. Ninety words were selected from the Malay-English translation norms established in Chapter 2. Following the recommendation of previous studies (Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014), the 90 words were selected from the full range of frequency bands to ensure that the test covered high frequency words that are most likely to be known by most Malay speakers, as well as low frequency words that are more likely to be known only by highly proficient Malay L1 and dominant speakers. Table 3.2 summarises the distribution of word stimuli across five frequency bands in Zipf values (van Heuven et al., 2014). From each frequency band, we sampled both easy (accuracy rate > 50%) and difficult word items (accuracy rate < 50%; based on the lexical decision accuracy data acquired from the Malay Lexicon Project, M. J. Yap et al., 2010). The final word list consisted of 46 nouns, 27 verbs, and 17 adjectives. Of these words, 60 were root words and 30 were words with circumfixes.

Table 3. 2

Distribution of word stimuli across frequency bands (in Zipf values)

Frequency band	Total number of words	Words with Acc_{LD} > .5			Words with Acc_{LD} < .5		
		n	*M*	*SD*	*n*	*M*	*SD*
Zipf < 3.0	21	7	.70	.11	14	.27	.14
3.0 ≤ Zipf < 3.5	25	8	.65	.10	17	.28	.13
3.5 ≤ Zipf < 4.0	20	7	.69	.12	13	.35	.14
4.0 ≤ Zipf < 5.0	20	15	.78	.16	5	.22	.20
Zipf > 5.0	4	4	.95	.03	-	-	-

Note. Acc_{LD}: Lexical decision accuracy rate obtained from M. J. Yap et al. (2010).

In addition to the 90 words, 90 pronounceable nonwords were also included in the LexMAL prototype to correct for response bias (e.g., participants answering “yes” to every stimulus to increase their scores). These nonwords were generated based on another set of 90 source words selected from the Malay-English translation norms gathered in Chapter 2 using the same selection criteria as for the word stimuli. A nonword generator, Pseudo (van Heuven, 2020) was employed to create nonwords (pseudowords) with legal letter combinations (bigrams and trigrams) in Malay. To achieve that, Pseudo randomly substituted one letter of the source words and checks the legality of the letter combinations within the nonword using bigrams and trigrams extracted from a corpus of 34,326 Malay words from the Malay Lexicon Project (M. J. Yap et al., 2010) and open-source spell checkers (Aspell[8] and Hunspell[9]). A set of 90 generated pseudowords were then later matched with the 90 word stimuli selected for the LexMAL prototype, in terms of their word length and orthographic neighborhood size, ts ≤ 0.32, ps > .75. The 90 nonwords were also checked against two Malay dictionaries, *Kamus Melayu-Inggeris Dewan* (Jasmani, 2012) and *Kamus Perdana* (S. H. Cheng & Lai, 2019) to ensure that these nonwords do not exist as real words in Malay. Finally, a LD1NN algorithm check (Keuleers & Brysbaert, 2011) was conducted on the combined list of words and pseudoword stimuli to verify that there was no inherent bias in terms of wordlikeness between the two stimuli sets (e.g., pseudowords appear to be too word-like or familiar compared to word stimuli), z = -0.95, p = .34. This ensured that vocabulary knowledge is needed for test takers to correctly identify words and pseudowords stimuli in LexMAL.

Task 2: Malay-English Bidirectional Translations. The Malay-English translation task consisted of 30 Malay nouns selected from the Malay-English translation norms established in Chapter 2. To avoid ceiling performance of Malay L1 speakers, translation stimuli with a moderate to high level of difficulty were chosen. The selection of word stimuli followed the criteria set out in Lemhöfer and Broersma's (2012) study, such that Malay (source) words with at least 50% translation error rates (including both omission and incorrect translations) and less than three possible English (target) translations were selected. The selected words were Malay nouns that could be translated into single-word English nouns. These criteria ensured that the Malay nouns selected for the task had a high difficulty level but were not too translation

[8] https://ftp.gnu.org/gnu/aspell/dict/0index.html (accessed in December 2020)
[9] https://github.com/titoBouzout/Dictionaries (accessed in December 2020)

ambiguous. No cognates or words from the LexMAL prototype were included in the stimuli. In total, 21 root words and 9 circumfixed words were selected, with a mean error rate of 70.00% (*SD* = 14.35), a mean number of possible translations of 1.83 (*SD* = 0.82), and a mean word frequency (Zipf value) of 3.67 (*SD* = 0.56, *min* = 2.95, *max* = 4.93).

Thirty English words were included in the English-Malay translation task. In total, 15 English words were taken from English-Malay translation norms (from Chapter 2) and a further set of 15 words with similar translation difficulty were selected from the English-Chinese translation norms (Wen & van Heuven, 2017a). Words from English-Chinese translation norms were included because there was a lack of potential translation stimuli with similar difficulty in the Malay-English translation norms. Overall, the stimuli from the English-Malay translation norms had a mean error rate of 73.81% (*SD* = 16.51), a mean number of possible translations of 1.53 (*SD* = 0.83), and a mean word frequency (Zipf value) of 3.68 (*SD* = 0.55, *min* = 1.89, *max* = 4.40). The 15 English words from Wen and van Heuven (2017a) had a mean error rate of 62.44% (*SD* = 13.00), a mean number of possible translations of 1.93 (*SD* = 0.70), and a mean word frequency (Zipf value) of 3.33 (*SD* = 0.67, *min* = 2.47, *max* = 4.58). Importantly, there was no significant difference between word frequencies (Zipf values) of words from both translation norms, *p* = .14. This ensured the difficulty level of the source words was matched despite being taken from different translation norms.

To ensure the appropriateness of test difficulty level of the translation tasks, a pilot test was conducted with 10 Malay L1 and 10 Malay L2 speakers. All items were translated correctly by at least one Malay L1 speaker. Neither floor nor ceiling effects were observed in the translation accuracy of the L1 (*M* = 51.50%, *SD* = 11.80%) and L2 (*M* = 32.67%, *SD* = 12.25%)

speakers. The final complete set of stimuli is available on the OSF repository (see link in Section 3.5).

Task 3: Malay-Mandarin Bidirectional Translations. A total of 30 Malay words were included in this task. Because there are no norms for Malay-Mandarin translation, 15 of the Malay words were selected from the Malay-English translation norms (from Chapter 2) and 15 words were selected from the English words of the English-Chinese translation norms (Wen & van Heuven, 2017a). Similar to the English-Malay translation task, words from English-Chinese translation norms were included to supplement the translation stimuli from Malay-English translation norms with similar translation difficulty. These English words were replaced with their Malay translation obtained from the *Kamus Dwibahasa* (Ibrahim, 2002) and the Oxford English-English-Malay Dictionary (Oxford University Press & Oxford Fajar, 2018). When an English word had more than one possible Malay translations, the Malay word that, according to *Kamus Perdana* (S. H. Cheng & Lai, 2019), had its dominant meaning matched with the dominant Mandarin translation (Wen & van Heuven, 2017a) was selected. No cognates were included and all words were nouns. The word frequency (Zipf value) of the Malay stimuli from the Malay-English (*M* = 3.66, *SD* = 0.53, *min* = 3.05, *max* = 4.67) and English-Chinese translation norms (*M* = 3.69, *SD* = 0.62, *min* = 2.95, *max* = 4.60) were matched, *p* = .88.

The Mandarin stimuli for the Mandarin-Malay translation task consisted of Mandarin translations of the 15 Malay words selected from the Malay-English translation norms (from Chapter 2), and 15 Mandarin dominant translations of 15 English words selected from the English-Chinese translation norms (Wen & van Heuven, 2017a). For Malay words that had more than one possible Mandarin translation, Mandarin words that were matched with the English dominant translations (from Chapter 2) and the dominant meaning of the Malay source words were chosen (based on *Kamus Perdana*; S. H. Cheng & Lai, 2019). Word frequency information for these Mandarin translations were obtained from Cai and Brysbaert (2010). Overall, the word frequency (Zipf values) for stimuli from the Malay-English (*M* = 3.85, *SD* = 0.76, *min* = 1.95, *max* = 4.73) and English-Chinese translation norms (*M* = 4.03, *SD* = 0.65, *min* = 2.43, *max* = 5.17) were matched, *p* = .50.

The translation stimuli were piloted with the same group of Malay L2 speakers who had participated in the Task 2 pilot test. No floor or ceiling effect was found (*M* = 46.17%, *SD* = 14.64%). However, two Mandarin (i.e., *炽热/bahang* and *心算/congak*) and three Malay items (i.e., *tikai/差别, komplot/阴谋* and *istilah/术语*) from the Mandarin-Malay and Malay-Mandarin translation tasks respectively were replaced with other words that matched the selection criteria mentioned above because they received no correct translation during the pilot test. The final set of word stimuli used for this task is available on the OSF repository (see link in Section 3.5).

Task 4: Malay Cloze Task. Twenty Malay cloze questions were randomly selected from several Malay sample examination papers that were designed for students of different education levels. Among the 20 questions selected, five easy questions (25%) were randomly sampled from the *Ujian Pencapaian Sekolah Rendah* paper (UPSR – the official examination taken by Malaysian students at primary sixth grade). The other 15 harder questions (75%) were randomly taken from the *Penilaian Tingkatan 3* (PT3 – the examination taken by Malaysian students at secondary third-form grade). The cloze questions involved a multiple-choice format that assessed vocabulary knowledge. The vocabulary items tested include nouns, verbs, and adjectives.

The difficulty level of the cloze questions was piloted using six Malay L1 and seven Malay L2 speakers. As expected, the L1 speakers performed at high accuracy with smaller variation (*M* = 90.83%, *SD* = 6.72%), whereas the L2 speakers scored lower with higher variability (*M* = 58.57%, *SD* = 15.29%).

Task 5: Self-ratings and Language Background Questionnaire. A language background questionnaire was created based on the Language History Questionnaire 3 (the same questionnaire as used in Chapter 2; P. Li et al., 2020). The questionnaire was used to acquire 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 the known languages. The questionnaire also asked for self-rated proficiency for Malay, English and Mandarin (Mandarin L1 participants only), using a scale from 1 (*very poor*) to 7 (*native-like*).

3.2.1.3 General Procedure

The present experiment was administered online using Qualtrics (https://www.qualtrics.com). Participants were instructed to complete all tasks without external aids (e.g., dictionary). The study was approved by the Ethics Committee in the School of Psychology at the University of Nottingham Malaysia. Written consent was obtained from participants before data collection started.

The study started with the LexMAL prototype. Participants were required to make unspeeded yes/no decision to every stimulus presented to them, one at a time. The words and nonwords were presented to all participants in the same randomized order. Care was taken to ensure that in the random order stimuli of the same type (i.e., word/nonword) did not appear in four consecutive trials. 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.

Next, participants completed the Malay-English translation task before the English-Malay translation task. Translation stimuli appeared one at a time on screen, and participants were required to enter the first translation that came to their mind. They could skip an item by indicating that they did not know the word or if they could not provide a translation. The Malay L2 speakers were presented with the Malay-Mandarin bidirectional translation tasks after completing the Malay-English bidirectional translations.

The Malay cloze task was presented after the translation tasks. Questions appeared on screen one at a time, and participants were required to select one correct answer out of four available choices. After that, the language background questionnaire was presented as the last part of the study. Participants were expected to spend about 45 minutes to complete all tasks.

3.2.2 Results

To ensure that participants processed the stimuli before responding (i.e., lexical access), we excluded those who responded quicker than 300ms for a significant number of trials (> 5%) because there does not seem to be sufficient time for the word stimuli to be processed. The data of two participants from the L2 group were excluded from data analysis because their response times in the LexMAL prototype were unusually fast. Additionally, a third participant was excluded from the data analysis because of the close-to-chance performance in the LexMAL prototype (18.33%; c.f. 0%, which indicates all words were incorrectly identified as nonwords, and all nonwords were incorrectly identified as real words).

Item assessment was conducted with the remaining data to examine the quality of all 90 word and 90 nonword items tested in the LexMAL prototype. The first subsection below reports results of the item assessment and describes the process of item selection for the final version of LexMAL. Subsequently, validity of the final LexMAL was evaluated by comparing LexMAL scores between the Malay L1 and L2 speakers. Lastly, convergent validity of LexMAL was examined via its correlations with the scores of other language tasks. Test reliability of LexMAL was computed using Cronbach's alpha.

3.2.2.1 Item Assessment

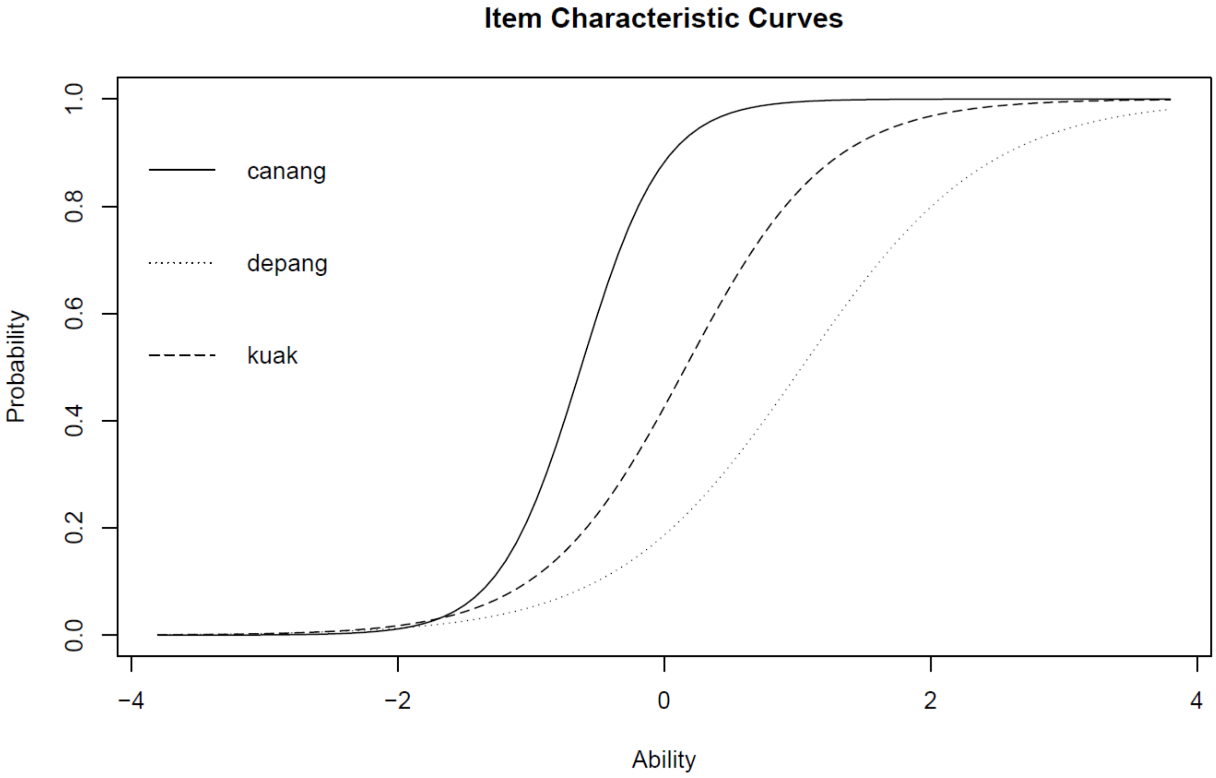
The approach used for the item assessment and selection of the final set of items for LexMAL was based on Wen et al. (2023). Behavioral data of the word and nonword items were assessed separately. Point-biserial correlations between the individual item responses and the overall test scores of participants were computed to assess predictiveness of each item to the overall test score. These correlations vary between -1.0 to +1.0. A positive point-biserial correlation indicates that good test performers (i.e., participants who obtained high overall scores) tend to identify the item correctly, when compared to weak test performers. In contrast, a negative point-biserial correlation reveals an atypical situation where the good test performers do less well on the item than the weak performers. Only items with positive point-biserial correlation were considered for the final version of LexMAL to achieve high test validity and reliability (Izura et al., 2014).

Out of the 90 words, 86 had positive correlations and four words (i.e., “*ambak*”, “*juru*”, “*memijakkan*”, “*sementara*”) yielded negative correlations (*r*s < -.116). Likewise, all but two (88/90) nonwords showed positive correlations. The two nonwords that had negative correlations were “*surindam*” (*r* = -.126) and “*abi*” (*r* = -.243). These six items with negative correlations were removed from subsequent analyses.

Next, the items in the LexMAL prototype were assessed in terms of their discriminatory power. An IRT analysis was conducted to examine how well each test item distinguishes speakers according to their Malay proficiency (Amenta et al., 2020; Brysbaert, 2013; Izura et al., 2014; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022). As a modern test theory method, IRT modelling estimates item parameters (e.g., item difficulty) independently of the specific group of participants that responded to the items (Paek & Cole, 2019). In the same vein, it evaluates a person's latent ability (e.g., vocabulary knowledge) independently of the specific set of items to which they responded to. IRT analysis provides a measure of the difficulty level and the discrimination power of each item. For this purpose, a two-parameter logistic model in the *ltm* R package (Rizopoulos, 2006) was used to assess word and nonword items separately. The IRT analysis represents the speakers’ ability range on the *x*-axis, and the probability to answer the item correctly on the *y*-axis. The difficulty level of an item was operationalised by the ability level of participants who have 50% chance to answer the item correctly (i.e., at 0.5 probability). On the other hand, the discrimination power, or how well an item can differentiate between speakers of different proficiency levels, was operationalised by the steepness of item response curve. The final set of the test items were chosen so that they span over the entire difficulty range, and have steep item response curves. Figure 3.1 presents the item characteristic curves for three word items of LexMAL. Based on the curves, “*depang*” was more difficult than “*canang*” and “*kuak*”, whereas “*canang*” had higher discrimination power compared to the other two words.

Figure 3. 1

Example of Item Characteristic Curves



The IRT analysis revealed three word items (“*mengehadkan*”, “*pemilihan*”, “*serta*”) with negative discrimination power, indicating that these items did not accurately discriminate between participants with high and low proficiency. Specifically, “*pemilihan*” and “*serta*” were rather easy words, hence all participants were able to identify the words. In contrast, “*mengehadkan*” was more consistently identified by participants with lower test scores, and missed by seven participants from the mid-to-high performance range. These three words (“*mengehadkan*”, “*pemilihan*”, “*serta*”) were excluded from the stimulus set. Subsequently, the remaining 83 words were ordered according to their difficulty level, from the lowest to the highest. Thirty difficulty groups were formed by grouping the ordered items into 23 groups of three items and seven groups of two items. Word items for the final LexMAL were selected by choosing two words with the highest discrimination power from each difficulty group (Amenta et al., 2020; Brysbaert, 2013; Izura et al., 2014; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022). No significant difference was observed in the discrimination power of words from the low (*M* = 2.43) and high (*M* = 2.06) difficulty groups (created by splitting around the mean difficulty level, -1.76), *t*(20.78) = 0.32, *p* = .75.

The IRT analysis of the nonwords revealed that all nonwords yielded discrimination power in the expected direction. Similar to the procedure used for the words, the 88 nonwords were ordered from the lowest to the highest difficulty level, and divided into 30 groups, in which 28 groups had three items and two groups had two items. The item with the highest discrimination power was selected from each difficulty group to form the final set of items for LexMAL. No significant difference was observed in the discrimination power of nonwords from the low (*M* = 1.85) and high (*M* = 2.06) difficulty groups (segregated by splitting around the mean difficulty level, -1.36), *t*(19.99) = -0.82, *p* = .42.

The above item selection procedure resulted in the most discriminative 60 word and 30 nonword items from the full range of difficulty levels. These final 90 items were selected for the final version of LexMAL. Table 3.3 summarises the lexical information of the selected items. No significant differences were observed in word length and orthographic neighbourhood size of word and nonword items, *t*s ≤ 0.75, *p*s ≥ .45.

Table 3. 3

Lexical Information of the Final Set of 60 Words and 30 Nonwords in LexMAL

Variable	Words		Nonwords	
	Mean	*SD*	*Mean*	*SD*
Number of letters	7.28	2.53	7.43	3.04
Orthographic neighbourhood	4.62	4.91	3.87	4.21
Word frequency (Zipf)	3.56	0.54	-	-

Note. Orthographic neighbourhood reported was Coltheart's N (Coltheart et al., 1977). It was computed using the *vwr* R package (Keuleers, 2011).

3.2.2.2 Discriminatory Power of Different Language Tasks

The normalised Ghent score was used for LexMAL scoring (following Wen et al., 2023). It ranges from -100% to 100%, with a negative score indicating a higher false alarm rate than correct word identification.

For the scoring of the responses in translation tasks, the Malay-English translations provided by the participants were checked against four Malay-English dictionaries: *Kamus Melayu-Inggeris Dewan* (Jasmani, 2012), *Kamus Perdana* (S. H. Cheng & Lai, 2019), *Kamus Dwibahasa* (Ibrahim, 2002), and the Oxford English-English-Malay Dictionary (Oxford University Press & Oxford Fajar, 2018). Likewise, the Malay-Mandarin translations were checked against four Malay-Mandarin dictionaries, namely *Kamus Perdana* (S. H. Cheng & Lai, 2019), Kamus Kembangan (Lai, 2018), *Kamus Cina-Melayu Dewan* (Jasmani, 2013), and the Chinese Malay English Dictionary (United Publishing House, 2019). The scoring criteria were the same as in Chapter 2. Correct translations with grammatical affixation that do not change the meaning of root words, such as the use of third person singular '-s' and plural '-s' in English, were collated to its root word and accepted as correct responses. Words with affixations that have a different word meaning or word class than the correct translations were classified as incorrect responses. Following scoring procedure from Lemhöfer and Broersma (2012), translations with spelling errors were classified as correct when errors differed by one letter from the correct translations and did not result in an existing word in the target language.

To illustrate the validity of LexMAL and other criterion measures in discriminating the vocabulary knowledge of Malay L1 and L2 speakers, independent *t*-tests were conducted to compare the performance between Malay L1 and L2 participants (see Table 3.4 for the average scores of each language group). As predicted, the Malay L1 group outperformed the Malay L2 group in all language tasks. Figure 3.2 summarises the distribution of the performance gap between the L1 and L2 participants for each language task. Large effect sizes (Cohen's *d*) were found for all language tasks, except for the Malay-English translation task which revealed a medium effect size between L1 and L2 speakers' performance. Specifically, the L1-L2 differences were larger for LexMAL and cloze test compared to that of translation tasks.

Table 3. 4

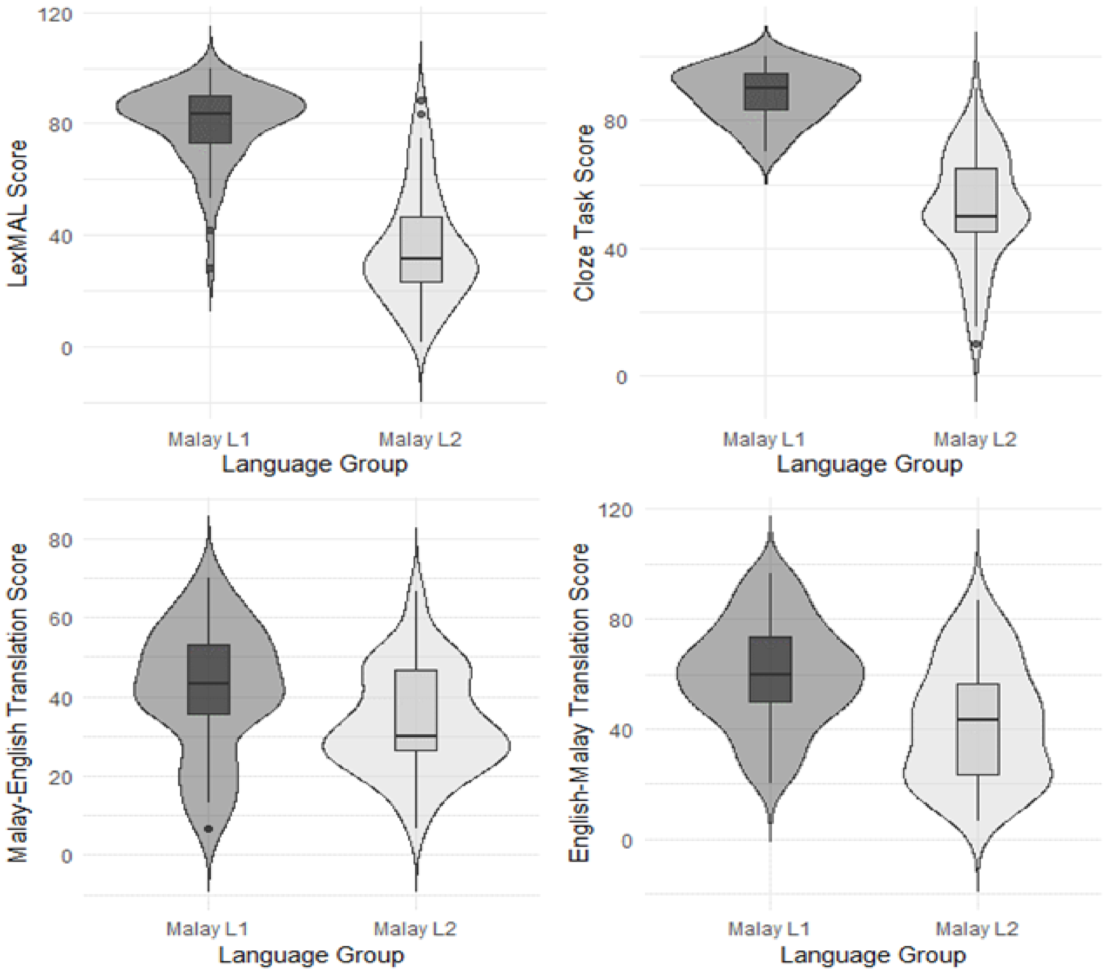
Test scores of all language tasks for both language groups in Experiment 1

Language Tasks	Malay L1 (*n* = 60)		Malay L2 (*n* = 57)		*t*	*df*	Cohen's *d*
	Mean	*SD*	*Mean*	*SD*			
LexMAL	90.04	6.88	67.75	10.04	13.95**	98.49	2.59
Malay-English Bidirectional Translation							
Malay-English	41.61	16.08	33.80	13.08	2.87*	115	0.53
English-Malay	59.83	18.97	41.93	21.33	4.80**	115	0.89
Combined	50.72	16.63	37.87	16.27	4.22**	115	0.78
Malay Cloze Test	88.33	8.32	52.63	17.35	14.08**	79.54	2.62

Note. * $p \leq .05$; ** $p < .001$.

Figure 3. 2

Distribution of Malay L1 and L2 speakers' test scores for all language tasks



3.2.2.3 Correlations of LexMAL with Other Language Tasks

To examine the validity of LexMAL as a lexical proficiency measure, correlational analyses were conducted to investigate the relationship between LexMAL and self-rated Malay proficiency with other vocabulary knowledge measures. Table 3.5 summarises the Pearson's correlation coefficients. LexMAL scores and self-ratings of all participants correlated positively and moderately with scores of the other language tasks. Although LexMAL scores and self-rated proficiency were strongly correlated, participants with identical self-rated proficiency varied considerably in their LexMAL scores (e.g., 95% *CI* [49.04, 81.29] at self-rated proficiency of 6 – *very good*, as demonstrated in Figure 3.3). Furthermore, LexMAL scores discriminated better between Malay L1 and L2 speakers, because Malay L1 speakers (e.g., 95% *CI* [74.48, 92.85] at self-rated proficiency of 6 – *very good*) systematically scored higher than L2 speakers (e.g., 95% *CI* [27.70, 65.63] at self-rated proficiency of 6 – *very good*) even when they rated their Malay proficiency at the same level.

Figure 3. 3

Correlation between self-rated Malay proficiency and LexMAL scores

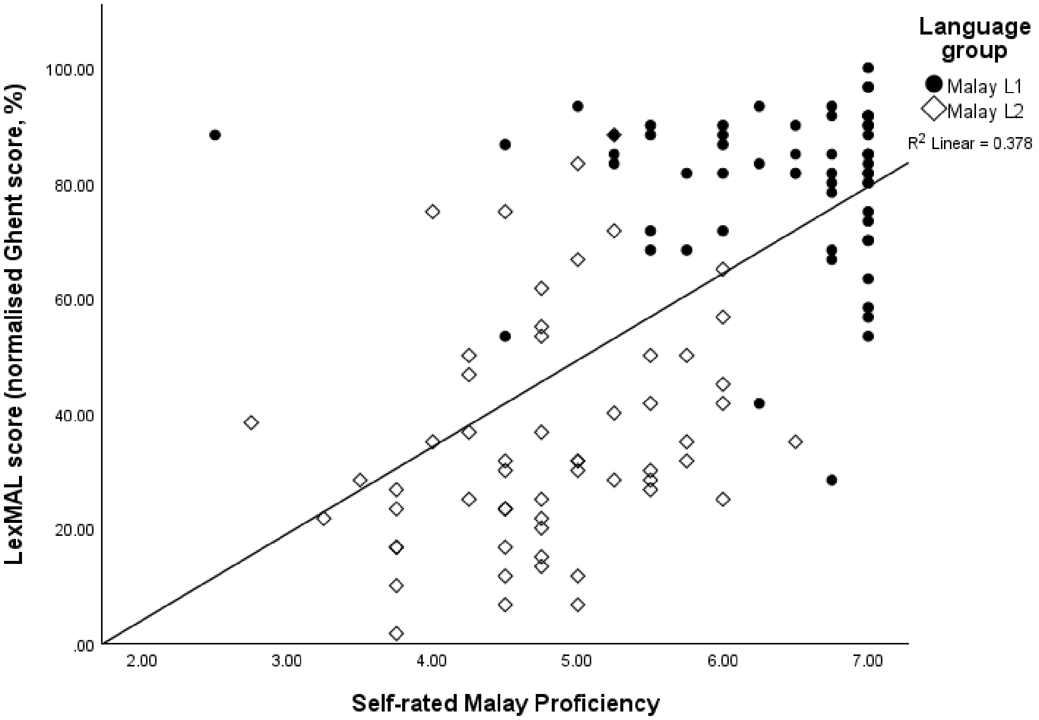


Table 3. 5

Correlations of LexMAL scores and self-ratings with other language tasks

Predictor	All participants (*N* = 117)				Malay L1 (*n* = 60)				Malay L2 (*n* = 57)					
	Lex	ME	EM	Malay cloze	Lex	ME	EM	Malay cloze	Lex	ME	EM	MC	CM	Malay cloze
LexMAL	1.00	.37***	.51***	.78***	1.00	.18	.20	.37**	1.00	.40**	.41**	.62***	.34*	.42***
SR														
Listening	.52***	.32***	.43***	.63***	-.08	.21	.11	-.03	.13	.22	.35*	.34*	.44***	.40**
Speaking	.63***	.31***	.41***	.64***	.04	.15	.08	-.05	.34*	.24	.34*	.44***	.41**	.43***
Reading	.55***	.21*	.33***	.59***	-.09	.07	-.04	-.07	.20	.04	.22	.28*	.30*	.28*
Writing	.57***	.26**	.37***	.59***	-.01	.10	.07	.02	.25	.16	.23	.34*	.35*	.25
Average	.62***	.30***	.42***	.66***	-.04	.15	.06	-.03	.28*	.20	.34*	.42***	.45***	.41**

Note. SR: Self-ratings. Lex: LexMAL.

ME: Malay-English translation.

EM: English-Malay translation.

MC: Malay-Mandarin translation.

CM: Mandarin-Malay translation.

The highest significant correlation in each column is bolded. * *p* < .05; ** *p* < .01, *** *p* ≤ .001.

To examine whether LexMAL scores outperformed self-ratings in terms of their correlation with other language tasks, Williams' (1959) *t*-tests were conducted to compare the correlation strengths using the SPSS code from Weaver and Wuensch (2013). Results indicate that LexMAL scores correlated better than average self-ratings with Malay cloze test scores, *t*(114) = -2.54, *p* = .01. No significant difference was found between LexMAL scores and average self-ratings for the correlations with Malay-English bidirectional translations, *t*s ≤ 1.28, *p*s ≥ .21. Furthermore, the correlation of LexMAL scores with Malay cloze test scores was significantly higher than its correlation with Malay-English translation, *t*(114) = -5.65, *p* < .001, and English-Malay translation scores, *t*(114) = -4.28, *p* < .001. Interestingly, when the correlational analyses were restricted to the Malay L1 group, self-ratings and LexMAL no longer correlated with translation accuracy, *p*s ≥ .13, but LexMAL scores still correlated significantly with Malay cloze test scores, *r*(58) = .37, *p* = .003.

For Malay L2 speakers, LexMAL scores correlated positively with all other language tasks, *r*s ≥ .34, *p*s < .05. Similarly, their average self-ratings also correlated positively with all other language tasks, *r*s ≥ .34, *p*s < .05, except for their Malay-English translation scores, *p* = .14. With respect to correlation strength, Williams' (1959) *t*-test indicated no significant difference between the correlations of LexMAL scores and average self-ratings with other language tasks, *t*s ≤ 1.58, *p*s ≥ .12. In other words, the correlation strength of LexMAL with (a) English-Malay translation; (b) Malay-Mandarin bidirectional translations; and (c) Malay cloze test scores were comparable to those of average self-ratings.

3.2.2.4 Reliability

Cronbach’s alpha returned a high reliability score for the final LexMAL at .94, .82 when the analysis was restricted to the Malay L1 group and .84 when the analysis was limited to the Malay L2 group.

3.2.3 Discussion

The 180-item LexMAL prototype was tested in Experiment 1 to select a final set of 90 items that have the highest discriminative power and span across a wide range of difficulty levels. In addition to self-ratings (cf. Amenta et al., 2020; Brysbaert, 2013; Izura et al., 2014; Salmela et al., 2021; Zhou & Li, 2022), bidirectional translation tasks and a cloze test were used as the external criterion measure to validate LexMAL (cf. Lemhöfer & Broersma, 2012; Wen et al., 2023).

As predicted, the Malay L1 speakers outperformed the L2 speakers on all language tasks. Specifically, the largest effect sizes were found for LexMAL and the cloze test, indicating that these two tests are the most sensitive at detecting L1-L2 proficiency differences. Furthermore, LexMAL scores positively correlated with translation and cloze test accuracies, providing evidence to support the validity of LexMAL as a Malay proficiency measure. In addition, the strong correlation between LexMAL scores and cloze test accuracy was significantly higher than that of self-ratings and cloze test accuracy, advocating LexMAL as an objective language measure that provides a better Malay proficiency estimate for bilingual speakers. Overall, the validity evidence of LexMAL is in-line with LexTALE (Lemhöfer & Broersma, 2010) and its extensions (Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022).

3.3 Experiment 2: Validation Study

Experiment 1 demonstrated some validity and reliability evidence of LexMAL as a Malay proficiency measure for Malay L1 and L2 speakers. Participants in Experiment 1, however, were presented with the 180-item LexMAL prototype. Because the items in the final LexMAL test were reduced to 90, it is important to replicate the reliability and validity of LexMAL. In this section, the 90-item final LexMAL was tested with another group of Malay L1 and L2 speakers.

3.3.1 Methods

3.3.1.1 Participants

The same recruitment criteria and general procedures from Experiment 1 were followed for this validation study. A total of 122 Malay L1 (N = 61, 15 males and 46 females) and L2 speakers (N = 61, 15 males and 46 females) were recruited. All participants were screened to confirm that they did not participate in Experiment 1. All but one Malay L1 speaker identified Malay as their L1 and dominant language (the exceptional participant acquired English as his L1 before the acquisition of Malay at the age of five, which later also became his dominant language). All Malay L2 speakers acquired their L1 (Mandarin) before Malay, except for three participants who reported simultaneous exposure to Mandarin and Malay since birth. Importantly, these three participants identified Mandarin as their dominant language, just as other participants from the same language group. Similar to Experiment 1, the Malay L1 speakers' self-ratings for Malay proficiency were significantly higher, and age of acquisition was significantly younger, than the L2 speakers, *t*s ≥ 12.11, *p*s < .001 (see Table 3.6 for speaker's language background summary).

Table 3. 6

Summary of participants' language background

Variable	Malay L1			Malay L2		
	Range	*Mean*	*SD*	Range	*Mean*	*SD*
Age (years)	19–35	23.15	4.21	20–38	25.70	4.81
Age of acquisition (years)						
Malay	0–5	0.13	0.67	0–7	4.51	1.63
English	0–10	4.15	2.64	0–9	4.28	2.13
Mandarin				0–4	0.51	1.06
Self-rated proficiency						
Malay	4.25–7.00	6.25	0.76	1.75–6.50	4.54	0.80
English	4.00–7.00	5.38	0.65	3.25–6.25	4.74	0.73
Mandarin				3.50–7.00	5.90	0.87

Note. Language background questionnaire measured self-rated proficiency on a 7-point scale (1 = *very poor*, 7 = *native-like*).

3.3.1.2 Stimuli and Procedure

The final 90-item LexMAL was used in Experiment 2. Other tasks included in Experiment 2 (translations, cloze task and questionnaire) were identical to those used in Experiment 1. The procedure was identical to Experiment 1. The study was approved by the Ethics Committee of the School of Psychology at the University of Nottingham Malaysia. All participants provided informed consent at the beginning of the study.

3.3.2 Results

The same set of statistical analyses as Experiment 1 was conducted to examine if the validity evidence of LexMAL from the Preparatory Study could be replicated after the removal of items with low discrimination power. To evaluate validity of the 90-item final LexMAL, independent *t*-tests were conducted to compare LexMAL scores between the two language groups. Additionally, correlational analyses were conducted to evaluate convergent validity of the final LexMAL with the scores of other language tasks. The test reliability was computed using Cronbach's alpha.

3.3.2.1 Discriminatory Power of Different Language Tasks

Table 3.7 summarises the average scores of participants across different language tasks. Overall, the participants' performance was comparable to that of Experiment 1, except that the Malay L1 speakers' mean LexMAL score was significantly lower than that of L1 speakers in Experiment 1, *t*(109.66) = 2.39, *p* = .02, *d* = 0.43. Nevertheless, similar to Experiment 1, the LexMAL scores of the L1 and L2 speaker groups still differed at a large effect size.

Table 3. 7

Test scores of all language tasks for both language groups in Experiment 2

Language Tasks	Malay L1 (*n* = 61)		Malay L2 (*n* = 61)		*t*	*df*	Cohen's *d*
	Mean	*SD*	*Mean*	*SD*			
LexMAL	86.46	9.45	67.42	10.39	10.59**	120	1.92
Malay-English Bidirectional Translation							
Malay-English	40.82	11.92	36.23	11.67	2.15*	120	0.39
English-Malay	55.85	14.05	44.43	15.95	4.20**	120	0.76
Combined	48.33	10.18	40.33	10.86	4.20**	120	0.76
Malay Cloze Test	86.48	8.77	51.23	17.55	14.03**	88.22	2.54

Note. * *p* ≤ .05; ** *p* < .001.

3.3.2.2 Correlations of LexMAL with Other Language Tasks

LexMAL scores correlated positively with the scores of all other language tasks and self-ratings, hence replicating the convergent validity of LexMAL in Experiment 1 (see Table 3.8 and Figure 3.4). In addition, as in Experiment 1, Williams' (1959) *t*-test was conducted to compare the correlation strengths of LexMAL scores and self-ratings with other language tasks using the

SPSS code from Weaver and Wuensch (2013). Results revealed that the correlation strength between LexMAL scores and cloze test scores was significantly higher than that of Malay-English translation, *t*(119) = 4.51, *p* < .001, and English-Malay translation, *t*(119) = 3.63, *p* < .001. There was no significant difference between the correlation strength of LexMAL scores and average self-ratings with all other language tasks, *t*s ≤ .78, *p*s ≥ .44.

As in Experiment 1, when the analysis was restricted to the L1 group, LexMAL scores correlated positively with cloze test scores. Intriguingly, unlike Experiment 1, self-ratings of the L1 group, but not their LexMAL scores correlated positively with their Malay-English translation scores. The average self-ratings also correlated with cloze test scores. In terms of correlation strength, there was no significant difference between the correlation of LexMAL and self-ratings with cloze test scores, *t*(58) = .84, *p* = .40.

For the Malay L2 group, LexMAL scores continued to correlate positively with Malay-Mandarin bidirectional translations and cloze test scores, whereas average self-ratings only correlated with the latter. With respect to correlation strength, Williams' (1959) *t*-test did not detect a significant difference between the correlation strengths of LexMAL scores and average self-ratings with cloze test scores, *t*(58) = .45, *p* = .65.

Reliability analysis revealed that the Cronbach's alpha for final LexMAL was .92. When the analysis was restricted to either Malay L1 or L2 group only, the Cronbach's alpha remained high at .85.

Figure 3. 4

Correlation between self-rated Malay proficiency and LexMAL scores in Experiment 2

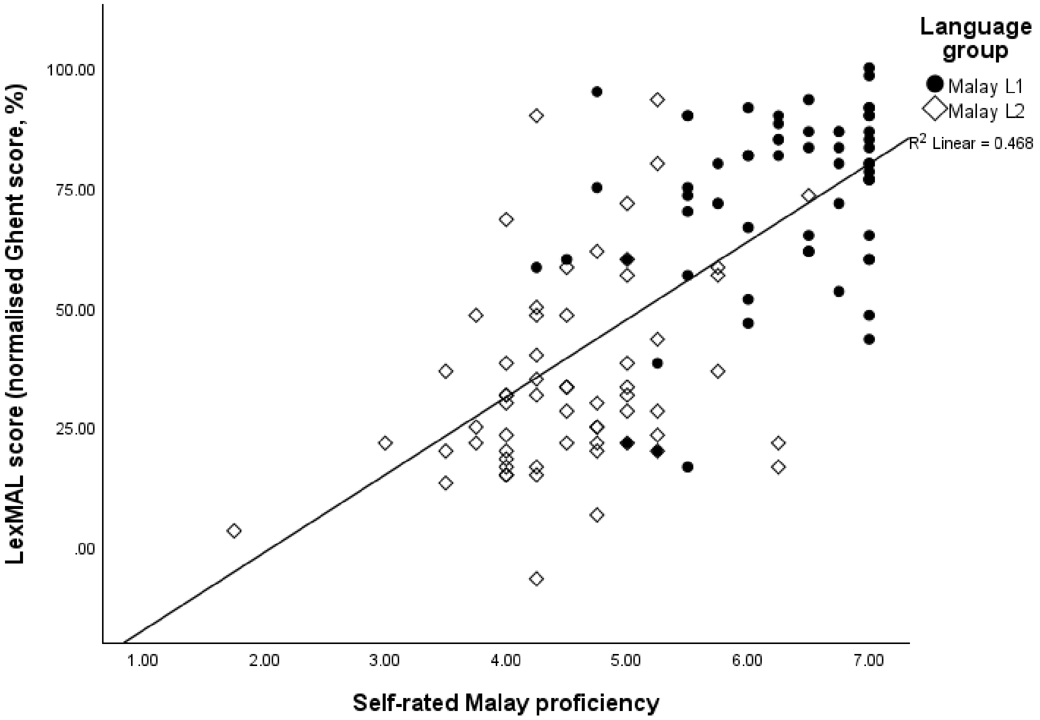


Table 3. 8

Correlations of LexMAL scores and self-ratings with other language tasks in Experiment 2

Predictor	All participants (*N* = 122)				Malay L1 (*n* = 61)				Malay L2 (*n* = 61)					
	Lex	ME	EM	Malay cloze	Lex	ME	EM	Malay cloze	Lex	ME	EM	MC	CM	Malay cloze
LexMAL	1.00	.28**	.39***	.69***	1.00	.15	.18	.41**	1.00	.25	.24	.31*	.34**	.29*
SR														
Listening	.63***	.26**	.35***	.69***	.18	.23	.10	.03	.28*	.14	.15	.12	.14	.35**
Speaking	.67***	.25**	.36***	.69***	.41**	.20	.19	.26*	.31*	.11	.15	.14	.22	.37**
Reading	.63***	.20*	.27**	.66***	.27*	.20	.04	.24	.32*	.01	.02	.33**	.02	.30*
Writing	.58***	.25**	.27**	.61***	.30*	.30*	.05	.40**	.28*	.02	.11	.19	.21	.23
Average	.68***	.26**	.34***	.72***	.36**	.28*	.12	.30*	.34**	.08	.12	.23	.17	.36**

Note. SR: Self-ratings.

Lex: LexMAL.

ME: Malay-English translation.

EM: English-Malay translation.

MC: Malay-Mandarin translation.

CM: Mandarin-Malay translation.

The highest significant correlation in each column is bolded. * *p* < .05; ** *p* < .01, *** *p* ≤ .001.

3.3.2.3 Discriminatory ability of LexMAL

In clinical settings, a receiver operator characteristic (ROC) curve analysis is frequently used to assess how well a diagnostic test can differentiate between two groups (e.g., people with or without a disease; Lalkhen & McCluskey, 2008; Read et al., 2016). Using a ROC curve analysis, Wen et al. (2023) proposed an optimum cut-off score that discriminates Mandarin L1 and L2 speakers at high sensitivity and specificity levels. To determine if LexMAL can distinguish between Malay L1 and L2 speakers, a ROC curve was plotted using the *pROC* R package (Robin et al., 2021).

Figures 3.5 and 3.6 presents the ROC curve for LexMAL plotted using data from Experiment 2. LexMAL's true positive rate (sensitivity) was plotted on the y-axis and false positive rate (1 – specificity) was plotted on the x-axis. The area under the ROC curve (AUC) measures LexMAL's ability to discriminate between L1 and L2 speakers' vocabulary scores, where an AUC of .5 indicates no discrimination ability, whereas an AUC of 1.0 indicates perfect discrimination (Hoo et al., 2017). The optimal cut-off point for LexMAL scores was also identified using point closest-to-(0, 1) corner method. The curve has an AUC of .892, suggesting that the LexMAL scores of Malay speakers correctly discriminated the L1 speakers from the L2 speakers at 89.2% of the time. An optimal cut-off point for LexMAL scores was identified at 59.2%, with the sensitivity and specificity of LexMAL being 86.9% and 82.0% respectively.

Figure 3. 5

ROC curve with data from Experiment 2

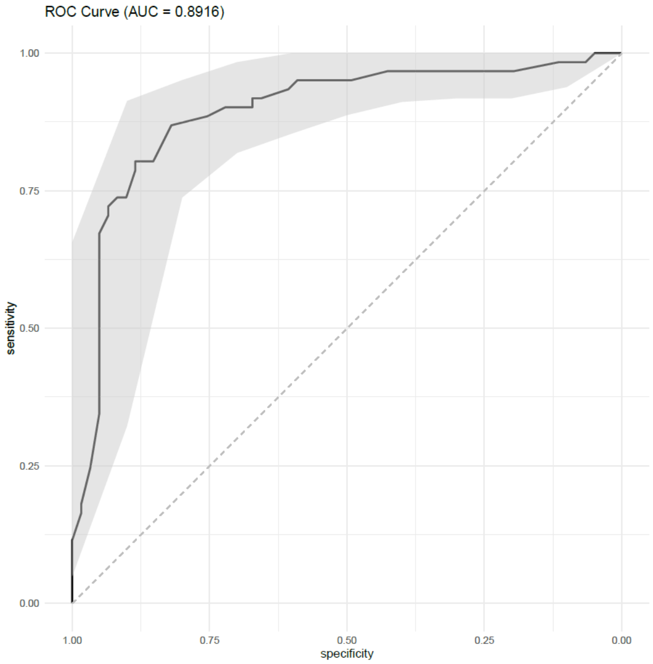
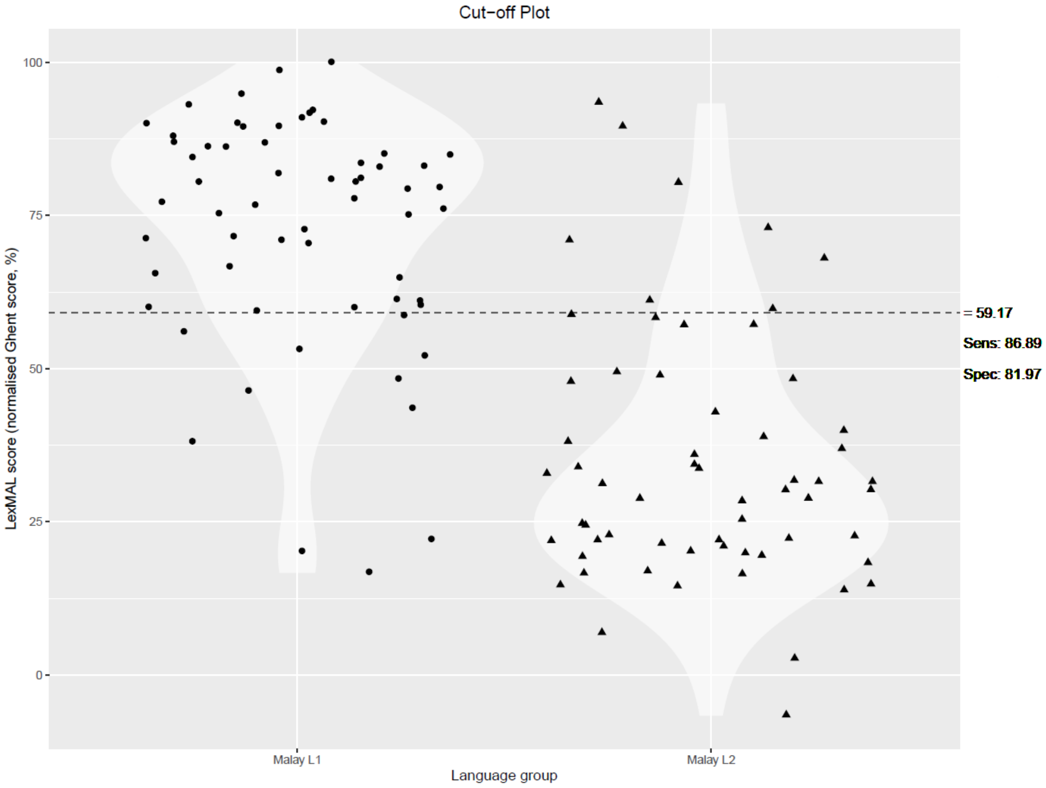


Figure 3. 6

Cut-off plot with data from Experiment 2



Note. Sens: Sensitivity. Spec: Specificity.

The same trend of LexMAL's discrimination ability, sensitivity, and specificity was observed, with a second ROC curve being plotted using data from both Experiments 1 and 2 (see Figures 3.7 and 3.8). The curve indicated that the Malay L1 speakers can be correctly discriminated from the L2 speakers 91.8% of the time. With the same cut-off score at 59.2%, the sensitivity and specificity of LexMAL were 86.4% and 86.0% respectively.

Figure 3. 7

ROC curve with data from Experiment 1 and 2

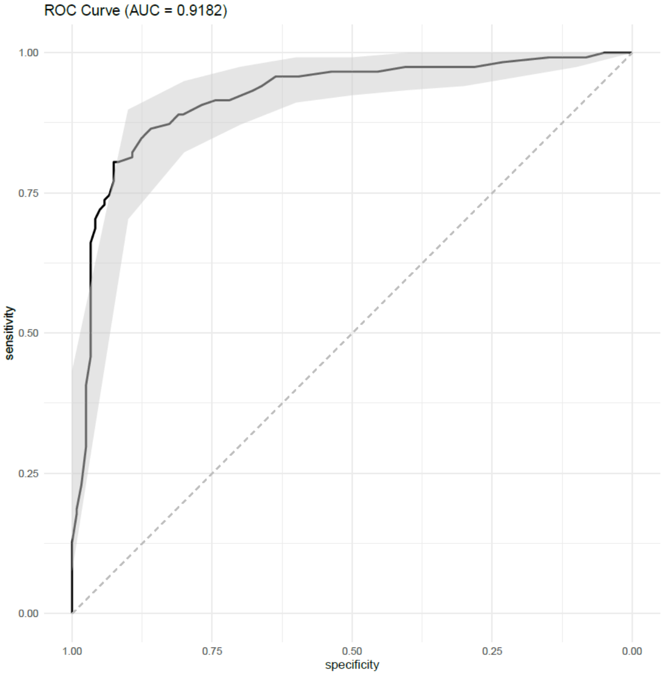
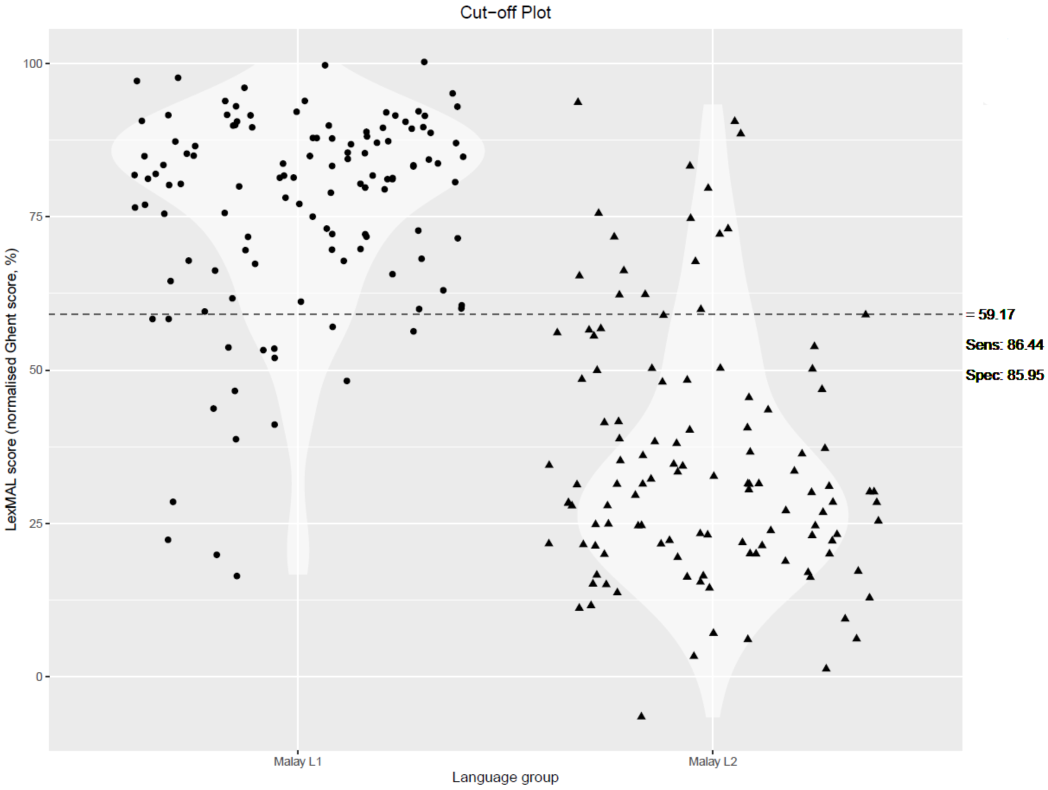


Figure 3. 8

Cut-off plot with data from Experiment 1 and 2



Note. Sens: Sensitivity. Spec: Specificity.

3.3.3 Discussion

Experiment 2 replicated the findings from Experiment 1, whereby the Malay L1 speakers consistently outperformed the L2 speakers at moderate-large effect sizes (Cohen's *d* ≥ 0.39). Similar to Experiment 1, LexMAL discriminated Malay L1 and L2 speakers reliably with a large effect size. Although the effect size in Experiment 2 was smaller compared to Experiment 1, the L1-L2 difference remains large. This suggests that discarding stimuli with lower discrimination power from the LexMAL prototype (in Experiment 1) does not reduce the discriminative sensitivity of the test. Nevertheless, the test might have become more difficult for highly proficient L1 speakers, given the lower LexMAL scores of L1 speakers in Experiment 2. Importantly, it remains useful in discriminating the Malay proficiency of L1 and L2 speakers, and this is further supported by the ROC curve analyses.

In concordance with Experiment 1, the convergent validity of LexMAL was demonstrated by the positive correlations between LexMAL score and other language task accuracies. LexMAL scores predicted the bilinguals' translation and cloze test performance. Furthermore, LexMAL scores and self-rated proficiency predicted cloze test performance equally well in Experiment 2. Both LexMAL scores and self-rated proficiency correlated strongly with cloze test accuracy, and there was no significant difference observed between the two correlation strengths.

3.4 General Discussion

The present study creates a quick valid Malay yes/no unspeeded vocabulary test to measure the proficiency of Malay L1 and L2 speakers. Following the development procedures used to develop LexTALE (Lemhöfer & Broersma, 2012) and its extensions (Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022), the LexMAL prototype was tested in Experiment 1. This prototype involved a larger stimulus set (180 stimuli) that was tested with two groups of speakers in Experiment 1. The final 90-item LexMAL was selected based on the results of Experiment 1 and the final LexMAL was tested in Experiment 2.

Due to a lack of freely available objective language proficiency test in Malay, past research has resorted to estimating Malay proficiency using self-reported measures such as order of language acquisition (e.g., L. W. Lee & Low, 2014; N. T. Yap et al., 2017) or self-ratings (e.g., Jalil et al., 2011; Rahman et al., 2018; Y. A. Rusli & Montgomery, 2020). However, considering most of the Malaysian Malay L2 speakers have a rather uniform age of Malay acquisition due to compulsory language education in school and their diverse language use and experience (Jin et al., 2013), individual differences in language proficiency of the bilingual or multilingual speakers can be difficult to assess based on just self-reported information. Hence, LexMAL as a freely available validated Malay proficiency test serves as a useful remedy that can objectively measure the proficiency of Malay L1 and L2 speakers for research purposes.

Just as LexTALE and its extensions, LexMAL is a yes/no unspeeded lexical decision task. Participants have to respond to one stimulus at a time by deciding yes or no depending on whether the letter string is a real word (Lemhöfer & Broersma, 2012). The validity of LexMAL was supported by the findings of both experiments reported in this chapter. Results showed that LexMAL scores can distinguish between Malay L1 and L2 speakers. Compared to other lextale extensions, a comparable effect size was found (see Table 3.9 for summary). Furthermore, no ceiling effect was observed for Malay L1 speakers and there was no floor effect observed for Malay L2 speakers. Thus, similar to Lextale_Fr (Brysbaert, 2013), Lextale_Esp (Izura et al., 2014), LEXTALE_CH (I. L. Chan & Chang, 2018), and LexCHI (Wen et al., 2023), LexMAL can be used with L1 and L2 speakers.

Table 3. 9

Comparisons of LexMAL scores with previous studies involving lextale extensions

Test	L1 speakers			L2 speakers			Cohen's *d*
	N	*Mean*	*SD*	*N*	*Mean*	*SD*	
LexMAL	60	90.0	6.9	57	67.8	10.0	2.59
Lextale_Fr	152	76.4	12.0	164	14.8	20.7	3.64
Lextale_Esp	91	89.8	11.0	123	19.8	29.8	3.11
LEXTALE_CH	49	73.2	9.8	15	25.8	19.8	2.91
LexITA	58	96.6	3.6	141	34.0	-	-
Lexize	117	89.4	16.6	159	39.3	27.6	-
LextPT	130	91.5	6.8	120	49.1	23.2	2.52
LexCHI	54	91.7	13.2	75	43.6	29.0	-

Note. All means are normalized Ghent score (Wen et al., 2023).

Malay L1 speakers in this study consistently outperformed the L2 Malay speakers in all language tests. However, it is worth noting that the translation tasks were not as sensitive as LexMAL and cloze test in discriminating speakers (as can be seen in Figure 3.2 there is large overlap in the translation scores of L1 and L2 speakers). The effect sizes of the performance differences were also smaller compared to LexMAL scores and cloze test scores (see Tables 3.4 and 3.7). Given that receptive (i.e., LexMAL and cloze tests) and productive tasks (i.e., translation tasks) involve linguistic knowledge and cognitive processes to different extents (Bialystok, 2001), participants' better performance in LexMAL and cloze test than translation tasks could also be attributed to the differences in difficulty level of the vocabulary knowledge and tasks involved. When assessing different aspects of vocabulary knowledge, recognition knowledge (e.g., identifying a word form given its meaning) is found to be easier and acquired earlier than recall knowledge (e.g., producing a word form given its meaning) (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004). Because the LexMAL and cloze tests do not require production of word form or meaning, participants are likely to rely more on recognition knowledge than recall knowledge, leading to better performance in the two tests. On the other hand, translation tasks are more challenging, even for the highly proficient L1 speakers, because it requires production of word forms. As a result, poorer translation performance is observed in both L1 and L2 speaker groups, compared to the LexMAL and cloze tests. Nonetheless, there remains a significant correlation between the test scores of recognition (LexMAL and cloze test) and recall (translation production), suggesting that they are measuring the same construct of vocabulary knowledge.

In terms of testing practicality, translation tasks are restricted to studies that involve bilinguals who speak the same language combination (e.g., English-Dutch, Lemhöfer & Broersma, 2012), and the scoring procedure is more time consuming compared to LexMAL and a cloze test (Webb, 2021). In summary, our findings indicate that LexMAL and cloze test are better options for studies seeking a quick and valid language proficiency measure of L1 and L2 speakers.

The convergent validity of LexMAL as a language proficiency measure was supported by significant correlations with translation and cloze test scores with moderate to large effect sizes (Cohen, 1988). In view of the high correlation between LexMAL scores and cloze test scores, one might easily assume that both LexMAL and cloze test might be equally reliable in measuring proficiency of Malay speakers. These two tests, however, are measuring different aspects of word knowledge (for detailed discussion, refer to Chapter 1.3). Specifically, a cloze test is a recognition test of collocations (knowledge of how words can be used together), whereas LexMAL is a test of form-meaning connections (i.e., vocabulary breadth). Correlations between these two tests were consistently found because knowledge of form-meaning connections to decode the meaning of words in sentences and word choices is necessary for correct responses to cloze questions (García & Cain, 2014; Gellert & Elbro, 2013; González-Fernández & Schmitt, 2020; Nation & Snowling, 1997; Schmitt, 2014). However, it is important to note that cloze tests adopt context-dependent testing (Read, 2000), in which grammatical knowledge is also involved to perform in the tests (Gellert & Elbro, 2013). In contrast, LexMAL presents words and nonwords in a de-contextualised manner (Amenta et al., 2020), which might provide a better estimate of construct distinct information about participants' word knowledge (Read, 2000).

LexMAL scores also strongly correlated with self-ratings, further supporting the validity of LexMAL as a language proficiency measure that corresponds to speakers' perception of own language proficiency. Specifically, when all participants were taken into consideration, participants who rated themselves with higher Malay proficiency tended to score higher on LexMAL. However, no significant correlation was found when the analysis was limited to the Malay L1 group in Experiment 1. The correlation between self-ratings of L1 speakers with their vocabulary test scores varied within and across previous studies. For instance, in the pilot studies of LEXTALE_CH (I. L. Chan & Chang, 2018) and LexCHI (Wen et al., 2023), the test scores showed weak and moderate correlations, respectively, with the self-rated proficiency of L1 speakers. However, these correlations were not present in the validation study. In these studies, L1 speakers usually showed smaller variance in their high vocabulary test scores when compared to L2 speakers (see Table 3.9 for a comparison between the *SD*s of L1 and L2 groups). It is likely that the homogeneity of their L1 vocabulary size as a group was one of the explanations for the negligible-weak correlation observed between the vocabulary test scores and self-rated proficiency (Brysbaert, 2013; I. L. Chan & Chang, 2018; Ferré & Brysbaert, 2017; Izura et al., 2014).

The subjectivity of self-ratings could also contribute to the lack of a correlation between the objective vocabulary measure and the subjective self-rated proficiency of L1 speakers. Unlike the L2 speakers who had both their self-ratings and LexMAL scores spread across the proficiency range, the L1 speakers showed greater variability in their self-ratings than their LexMAL scores (see Figure 3.3). When inspecting the LexMAL performance of Malay L1 and L2 speakers who gave themselves the same rating (e.g., 5/*good* – 6/*very good* in Figure 3.3), the majority of the Malay L1 speakers appeared to score higher than the L2 speakers. This is possibly due to the difference in reference group used by the Malay L1 and L2 speakers when rating their language proficiency. For instance, Brysbaert (2013) reported that Lextale_Fr participants from the L1 group tended to be stricter in self-ratings because they compared their language ability to other highly proficient L1 speakers. In contrast, the L2 speakers were more lenient because they compared their proficiency to other relatively less proficient L2 speakers. Hence, it is likely for the L1 speakers to be more proficient in the language compared to the L2 speakers even though they might provide the same rating (in line with our observation in Figure 3.3). Importantly, LexMAL scores, when compared to self-rated proficiency, correlated better with cloze task performance. Taken together, objective language proficiency measures like LexMAL are more sensitive in discriminating individual differences among speakers of a wide proficiency range, supporting the notion that objective language testing provides better estimates of language proficiency than subjective self-ratings (Khare et al., 2013; Lemhöfer & Broersma, 2012; Tomoschuk et al., 2019; Wen & van Heuven, 2017a).

Finally, the internal reliability analyses revealed that LexMAL is highly reliable in measuring the vocabulary size of Malay speakers. Due to the larger number of stimuli, it is not surprising that LexMAL's reliability is higher than that of LexTALE (a = .81, Lemhöfer &

Broersma, 2012). Such high reliability is also seen in other lextale extensions that have larger stimuli sets (Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022).

The ROC curve of LexMAL also suggests that the LexMAL score is a very good classifier of Malay proficiency in terms of Malay L1 and L2 speakers. Because Malay-English bilingual speakers in Malaysia use both languages in a variety of daily contexts from a very young age, it can be challenging for them to self-evaluate their L1 and L2 proficiencies and to indicate their order of acquisition and dominance. This is reflected in the less consistent prediction of self-rated proficiency on language task performance compared to LexMAL in the present study. For example, the correlations between LexMAL scores and cloze test accuracy for both L1 and L2 speakers were consistent across both experiments. Contrastively, the correlation between self-rated proficiency and cloze test accuracy was only consistent for L2 speakers across experiments, whereas for L1 speakers it correlated with cloze test accuracy only in Experiment 2 but not in Experiment 1. Hence, it appears that for bilingual populations in which people use two languages frequently from an early age and share a similar age of acquisition or order of language acquisition, an objective language proficiency measure like LexMAL provides a better estimate of language proficiency than self-rated proficiency.

In addition, LexMAL can also be used as a screening test to decide if a Malay-speaking bilingual has Malay proficiency resembling that of an L1 or L2 speaker. Nevertheless, unlike LexTALE and its Dutch and German parallel versions that were constructed to be as close to the English version as possible (Lemhöfer & Broersma, 2012), the lextale-inspired tests are created separately to measure bilinguals’ proficiency in one particular language (e.g., LexMAL for Malay proficiency; see also Section 1.3.2.3 for the list of lextale-inspired tests). Although the equivalence of the Dutch and German versions with LexTALE has yet to be tested, they offer a potential solution to compare the proficiency of various languages spoken by a bilingual or multilingual. In contrast, because the difficulty level of other lextale-inspired tests differs from that of LexTALE, direct inferences about the relationship between languages cannot be drawn by comparing the test scores (e.g., comparing LexTALE and LexMAL scores to infer which language is relatively stronger in a Malay-English bilingual). Nevertheless, the lextale-inspired tests provide a good foundation for future parallel test development to enable cross-linguistic comparison.

It is also important to note that LexMAL is designed to estimate Malay proficiency base on the vocabulary size measured. Despite its usefulness in research that seeks practical and objective proficiency measure, the present study does not provide direct evidence for how recognising the context independent LexMAL items measure written vocabulary knowledge (e.g., the vocabulary knowledge required for word recognition and recall). Therefore, future research is needed to pinpoint the extent to which the test measures form-meaning knowledge. Without such evidence, researchers should be cautious when LexMAL scores or scores of any LexTALE and lextale extensions are used as a reference. The next chapter addresses this gap by investigating the contribution of bilinguals' vocabulary knowledge, specifically knowledge of form and meaning, in yes/no vocabulary tests.

3.5 Conclusion

The present study described the development of LexMAL, a quick lexical test for estimating language proficiency in Malay. The validity and reliability of LexMAL as a Malay language proficiency measure was demonstrated, with no ceiling effect observed for the L1 speakers and no floor effect for L2 speakers. As far as we are aware, LexMAL is the first Malay lexical test that can reliably measure the proficiency of L1 and L2 speakers. LexMAL is useful for researchers in, for example, linguistics, psychology, and education that require a quick (less than 5 minutes), practical and objective proficiency measure. LexMAL is appropriate for assessing the Malay proficiency of speakers of Malay-Mandarin-English combinations because it has been validated with bilingual speakers of these language combinations. However, if the test is to be administered to bilinguals who speak different language combinations (e.g., Tamil speakers), the use of LexMAL should still call for cautious. For instance, test users should make sure that test takers' vocabulary knowledge of other languages does not contaminate their LexMAL performance by screening whether the words and nonwords in the test coexist as cognates or a different real word in those languages. LexMAL can be taken online at https://www.lexmal.org/, or a paper and pencil version of LexMAL can be downloaded from https://osf.io/8y4ft/.

Chapter 4

The Role of Vocabulary Knowledge in Answering LexMAL

In the last chapter, LexMAL was shown to reliably and validly measure bilinguals' language proficiency based on vocabulary knowledge. However, due to the lack of direct demonstration of word knowledge in the yes/no vocabulary test, it is unclear how bilinguals' vocabulary knowledge contributes to their language proficiency scores in LexMAL. This chapter investigates the contribution of bilinguals' form-meaning vocabulary knowledge to their item accuracy on LexMAL. Alongside LexMAL, four form-meaning vocabulary tests were developed to assess bilinguals' knowledge of meaning recognition, form recognition, meaning recall, and form recall. Malay L1 and L2 speakers' word knowledge on the 60 LexMAL word items were tested using these four tests. The statistical findings revealed that LexMAL score, as a measure of language proficiency, can be predicted by meaning recognition, form recognition, and meaning recall of word knowledge, further demonstrating the role of form-meaning vocabulary knowledge in performing yes/no vocabulary tests. Study reported in this chapter has been written up in a manuscript and submitted to Bilingualism: Language and Cognition for review:

Lee, S. T., van Heuven, W. J. B., Price, J. M., & Leong, C. X. R. (under review). Assessing Bilingual Language Proficiency with Yes/No Vocabulary Test: The Role of Form-Meaning Vocabulary Knowledge.

4.1 Introduction

Yes/no vocabulary tests such as the Lexical Test for Advanced Learners of English (LexTALE, Lemhöfer & Broersma, 2012) and Lexical Test for Malay Speakers (LexMAL) have been used to estimate bilinguals' language proficiency in psycholinguistic research. Positive correlations found between the yes/no vocabulary test scores and other language proficiency measures, such as Quick Placement Test (Lemhöfer & Broersma, 2012; Masrai, 2022) and translation tasks (Lemhöfer & Broersma, 2012; Wen et al., 2023; see also Chapter 3), had provided evidence for the convergent validity of these tests as language proficiency measures. These tests are freely available and time-efficient, taking less than five minutes to complete and allowing a large number of words to be tested in a short period of time. As demonstrated in Chapter 3, yes/no vocabulary tests can also be used to discriminate L1 and L2 speakers by grouping test takers into higher and lower proficiency groups based on their scores (in-line with previous studies e.g., Brysbaert, 2013; Izura et al., 2014; 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 validity of yes/no vocabulary tests has been consistently demonstrated in past studies (e.g., Lemhöfer & Broersma, 2012; Masrai, 2022; Wen et al., 2023; X. Zhang et al., 2020; see also Chapter 3), it is unclear precisely which aspects of test takers' vocabulary knowledge are assessed in these tests, making meaningful score interpretation problematic. A lexical decision format is used in the yes/no vocabulary tests, where test takers are required to decide whether the letter strings presented are real words. Subsequently, a "yes" response in the tests may reflect word knowledge that ranges from being able to recognise 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 recognise and produce the word forms or meanings when they correctly indicate a "yes" response in the yes/no vocabulary test. To this end, this chapter aimed to investigate the role of vocabulary knowledge to bilinguals' performance in a yes/no vocabulary test.

Most vocabulary tests to-date (e.g., LexTALE: Lemhöfer & 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 or breadth) at specific mastery level of form-meaning knowledge. Interpretation of the vocabulary test scores depends on the test format and 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' form recognition at the first five 1000-word frequency levels from the British National Corpus/Corpus of Contemporary American English (Nation, 2012a). Instead of assessing form recognition, the Vocabulary Size Test (Nation & Beglar, 2007) is designed to provide an estimate of English L1 and L2 speakers' overall receptive vocabulary size through meaning recognition. These vocabulary tests employ a multiple-choice format, and are widely used in language classrooms because it is found to reliably predict reading ability (e.g., Laufer & Aviad-Levitzky, 2017). However, there are drawbacks that 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 minutes 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 it with the knowledge of 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 yes/no vocabulary tests (e.g., LexTALE: Lemhöfer & Broersma, 2012; V_YesNo: Meara & Miralpeix, 2016; LexMAL: see Chapter 3), on the other hand, assess vocabulary knowledge as a distinct construct. Using a quick unspeeded lexical decision format, vocabulary knowledge is tested in a de-contextualised 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 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 tests present words and nonwords one at a time, and test takers are required to respond "yes" or "no" to indicate whether the letter string presented is a real English word. 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).

The yes/no vocabulary test format however, despite its simplicity, is not without flaws. In contrast to meaning recognition tests that assess knowledge of recognising word meaning given its word form (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, different interpretations of the scores have been proposed due to the vaguely defined target word knowledge aspect. Schmitt (2010) and X. Zhang et al. (2020), for example, proposed that correct responses in yes/no vocabulary tests require meaning recall knowledge before identity of the letter strings could 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 research 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 score and form-meaning knowledge to justify its score interpretation.

The present study aimed at filling this gap by investigating the relationships 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. We examined whether bilinguals' knowledge of form-meaning connections, namely meaning recognition, form recognition, meaning recall and form recall can affect their accuracy in a yes/no vocabulary test. In Chapter 3, LexMAL was shown to validly estimate language proficiency of Malay L1 and L2 speakers, and reliably discriminate the language proficiency of L1 and L2 speakers. As a continuation from the validation study, the present study presented LexMAL to Malay L1 and L2 speakers alongside four newly developed vocabulary tests that assess form-meaning knowledge to various degrees. In order to understand the impact of individual word knowledge on bilinguals' performance in the yes/no vocabulary test, the same set of words was tested across the four form-meaning vocabulary tests (following González-

Fernández & Schmitt, 2020; González-Fernández, 2022; McLean et al., 2020). At the test level, Malay L1 and L2 speakers' scores from the four form-meaning vocabulary tests were examined as predictors for LexMAL score (see Chapter 3) 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, item accuracy of each target word was compared across the vocabulary tests to evaluate the contribution of form-meaning knowledge to LexMAL accuracy.

The presentation order of the vocabulary tests was decided based on the difficulty hierarchy of form-meaning knowledge, progressing from the most difficult to the easiest (see Laufer & Goldstein, 2004, González-Fernández & Schmitt, 2020, and McLean et al., 2020 for similar approach). LexMAL (see Chapter 3) was first presented, followed by the more difficult form-meaning tests, which were the form recall test (write a word based on the meaning provided) and meaning recall test (explain the meaning of the word provided), and finally the easier counterparts, namely the form recognition test (select the correct word option based on the meaning provided) and meaning recognition test (select the correct meaning option based on the word provided). This approach ensured word exposure in the earlier vocabulary tests to not affect participants' responses in the later tests (Laufer & Goldstein, 2004; González-Fernández & Schmitt, 2020; McLean et al., 2020; Nation, 2013; Nation & Webb, 2011; Schmitt, 2010). Across the vocabulary tests, we expected Malay L1 speakers to score higher than the L2 speakers (as informed by Chapter 3, which is also in agreement with Bialystok et al., 2008; Fernandes et al., 2007; 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.

4.2 Method

4.2.1 Participants

To calculate for the required sample size, the weakest correlation found between LexMAL and cloze test scores in Experiment 1 of the LexMAL study was employed (see Table 3.5). The correlation was specifically chosen because it was consistent for both L1 and L2 speakers. The alpha level for the a prior power analysis was adjusted to .0125 (i.e., .05/4) using Bonferroni correction (as proposed by Vickerstaff et al., 2019) because the present study aimed to investigate the relationship between LexMAL scores and four form-meaning test scores. As a result, the a priori power analysis conducted using G*Power (Faul et al., 2009) indicated that 112 participants (56 in each language group) were required to obtain a .80 power to detect a medium effect size of .40 at the standard .0125 alpha error probability. One hundred and sixty Malay speakers (80 Malay L1 speakers, 70 females; 80 Malay L2 speakers, 65 females) participated in the study. All participants were included in the final analyses. A slightly larger sample size than recommended was recruited to accommodate for potential challenges in online research such as incomplete responses or participant dropouts. All participants were students or graduates of tertiary education and had a minimum “Pass (C)” qualification for the *Bahasa Melayu* (Malay) subject in Malaysian national high school examination (commonly known as the *Sijil Pelajaran Malaysia*, SPM). 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 4.1 for the summary of participants’ language background). They received monetary compensation for their participation.

Table 4. 1

Summary of participants' language background

Variable	Malay L1		Malay L2	
	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*).

4.2.2 Instruments

The present study comprised of 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 Questionnaire 3 (P. Li et al., 2020) was also presented to obtain information about participants' language background and experience.

4.2.2.1 Target words

The 60 Malay words from LexMAL (see Chapter 3) consisted of 31 nouns (22 root words, 9 words with "*pe-…-an*" circumfix[10]), 17 verbs (7 root words and 10 words with "*me-…-*

[10] Malay is a morphologically complex language where new words can be formed via rule-based affixation (M. J. Yap et al., 2010). For instance, a noun (e.g., "*penggelapan*/embezzlement") can be formed by adding a noun

kan" circumfix), and 12 adjectives. These words were a combination of high and low frequency words carefully chosen to assess both highly proficient and less proficient Malay speakers with good discrimination power (see Section 3.2.2.1 for the item selection process). The distribution of word stimuli across five frequency bands in Zipf values (van Heuven et al., 2014) is summarised in Table 4.2.

Table 4. 2

Distribution of target words across frequency bands (in Zipf values)

Frequency band	Total number of words		
	Noun	Verb	Adjective
Zipf < 3.0	5	4	3
3.0 ≤ Zipf < 3.5	11	4	3
3.5 ≤ Zipf < 4.0	8	5	1
4.0 ≤ Zipf < 5.0	7	4	5

Note. Word frequency of LexMAL items were obtained from M. J. Yap et al. (2010) and converted to Zipf values (van Heuven et al., 2014) for a more intuitive interpretation. The tipping point from low frequency to high frequency words is between 3.5 to 4 (van Heuven et al., 2014).

Only half of the 60 target words from LexMAL were presented for each of the subsequent vocabulary tests. Two wordlists (A and B) with matched word frequency and length (*t*s ≤ 0.50, *p*s ≥ .62) were created from the 60 target words (see Table 4.3 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

circumfix "*peN-…-an*" to an adjective "*gelap*/dark". In a similar way, a verb (e.g., "*menggelapkan*/darken") can be formed by adding a verb circumfix "*meN-…-an*" to the word "*gelap*/dark". These words with circumfixes can have different meanings and forms compared to the root words, and therefore were also tested in LexMAL (see Chapter 3).

tests, and the other wordlist for the Meaning Recall and Meaning Recognition tests[11]. 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 M. J. 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 seeing only the target words. Each vocabulary test (except LexMAL) presented 10 novel filler items in addition to the target words from wordlist A or B. Participants saw each filler item only once throughout the study. The target words and filler items were matched in terms of word frequency (Zipf value) and word length, *t*s ≤ .01, *p*s ≥ .93.

Table 4. 3

Distribution of lexical characteristics across wordlists

Word class	*N*	Wordlist A Word frequency	Word length	*N*	Wordlist B Word frequency	Word length
Noun	16	3.55 (0.57)	7.06 (2.11)	15	3.58 (0.55)	7.13 (2.77)
Verb	8	3.49 (0.52)	8.13 (2.95)	9	3.54 (0.53)	9.11 (2.93)
Adjective	6	3.52 (0.58)	6.00 (1.26)	6	3.73 (0.67)	5.67 (0.82)

Note. Word frequency was in Zipf value (van Heuven et al., 2014).

4.2.2.2 Vocabulary Test 1: Lexical Test for Malay Speakers (LexMAL)

LexMAL is an unspeeded yes/no vocabulary test designed to estimate the Malay proficiency of L1 and L2 speakers (see Chapter 3 for its development and validation). It contains

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

a total of 90 items (60 words and 30 nonwords) and participants were required to indicate if letter strings are existing Malay words by responding “yes” or “no”.

Scoring. LexMAL score (normalized Ghent score, see equation below; Wen et al., 2023) was computed by summing up the number of correctly identified word stimuli and penalizes the score base on guessing by the participant (“yes” responses for nonword stimuli, i.e., false alarms).

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

4.2.2.3 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* (Ibrahim, 2002). Because 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. For this purpose, words from the same frequency band[12], 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***

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

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 restrict the correct responses to the target words (excluding other words with similar meanings but different spellings), 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).

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, i.e., 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, i.e., number of letters only, *M* = 15.42%, *SD* = 25.25%, *t*(113.48) = 2.43, *p* = 0.02. Because 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 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 spelled correctly were marked as correct. The percentage of correct responses was used to compute the Form Recall score.

4.2.2.4 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, and participants were required to type the meaning of the target word in any languages 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. For instance, when considering the target word "*perenggan*/paragraph", an English translation such as "passage" that does not accurately convey the meaning of "paragraph", or an overly general description, like "*berkaitan dengan karangan*/related to essay", would be scored as incorrect. 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%)[13] with linguistics background were trained and scored

[13] The Malay L2 scorer also used Mandarin as L1. According to the ROC curve analysis in the previous chapter, the cut-off score for LexMAL to distinguish between L1 and L2 speakers was 59.2%. A LexMAL score of 50% is well

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 .94, whereas L2 responses were scored with 91.2% agreement and a Cohen's kappa of .81.

4.2.2.5 Vocabulary Test 4: Form Recognition

The Form Recognition test assesses the ability of recognizing 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 asked to choose the target word form that matches each definition. The target words were presented with three foils. In accordance with Nation (2012b) 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, 2012b). 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.

above the mean score of the L2 speakers in the present study (see Table 4.5). This scorer was tasked to score the Mandarin/Malay/English responses collected from the Malay L2 speakers.

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

4.2.2.6 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 (2012b), 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.

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

4.2.2.7 Language Background Questionnaire

A language background questionnaire based on the Language History Questionnaire 3 (P. Li et al., 2020) was used to acquire 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 of Malay, English and Mandarin (Mandarin L1 participants only), using a scale of 1 (*very poor*) to 7 (*native-like*).

4.2.3 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. Written consent was acquired from participants before data collection started.

The study started with LexMAL, in which the participants were required to make yes/no decision 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 link was not revealed to the participants. No feedback was provided to the participants, so that the unknown words remained unknown to them. Furthermore, test scores were not disclosed to the participants, maintaining a level of unawareness regarding their actual performance.

After LexMAL, a non-language filler task with 10 items adapted from Raven's progressive matrices task (Raven, 2000) was presented. Aiming to minimise potential interference from LexMAL in the subsequent vocabulary tests, this filler task presented shapes in a 3 by 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 fit 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 Form Recall test, followed by 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.

4.3 Results

Participants' data was first screened for unusually fast responses (less than 300ms for more than 5% of the trials) on the closed-ended vocabulary tests (i.e., LexMAL, Form Recognition and Meaning Recognition tests). No data was excluded at this stage as all response time fell within the normal limit. The mean total duration for the participants to complete each vocabulary test is summarised in Table 4.4. Participants' mean test scores are summarised in Table 4.5. The mean LexMAL scores of both L1 and L2 speakers were lower compared to that of in Chapter 3. Nevertheless, the mean LexMAL scores in the present study were still well above the cut-off score recommended in Chapter 3.

Table 4. 4

Mean total duration for vocabulary test completion

Vocabulary test	*N* of items	Mean total duration in minute (*SD*)
LexMAL	90	5.31 (3.00)
Form Recall	40	36.17 (17.28)
Meaning Recall	40	14.92 (10.18)
Form Recognition	40	5.26 (2.39)
Meaning Recognition	40	6.05 (2.72)

Table 4. 5

Means and standard deviations (in percentage) of accuracy for each vocabulary test

Vocabulary test	Language group					
	Malay L1 (*N* = 80)			Malay L2 (*N* = 80)		
	M	*SD*	Range	*M*	*SD*	Range
LexMAL	74.12	19.45	11.67-100.00	34.15	21.78	0.00-96.67
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

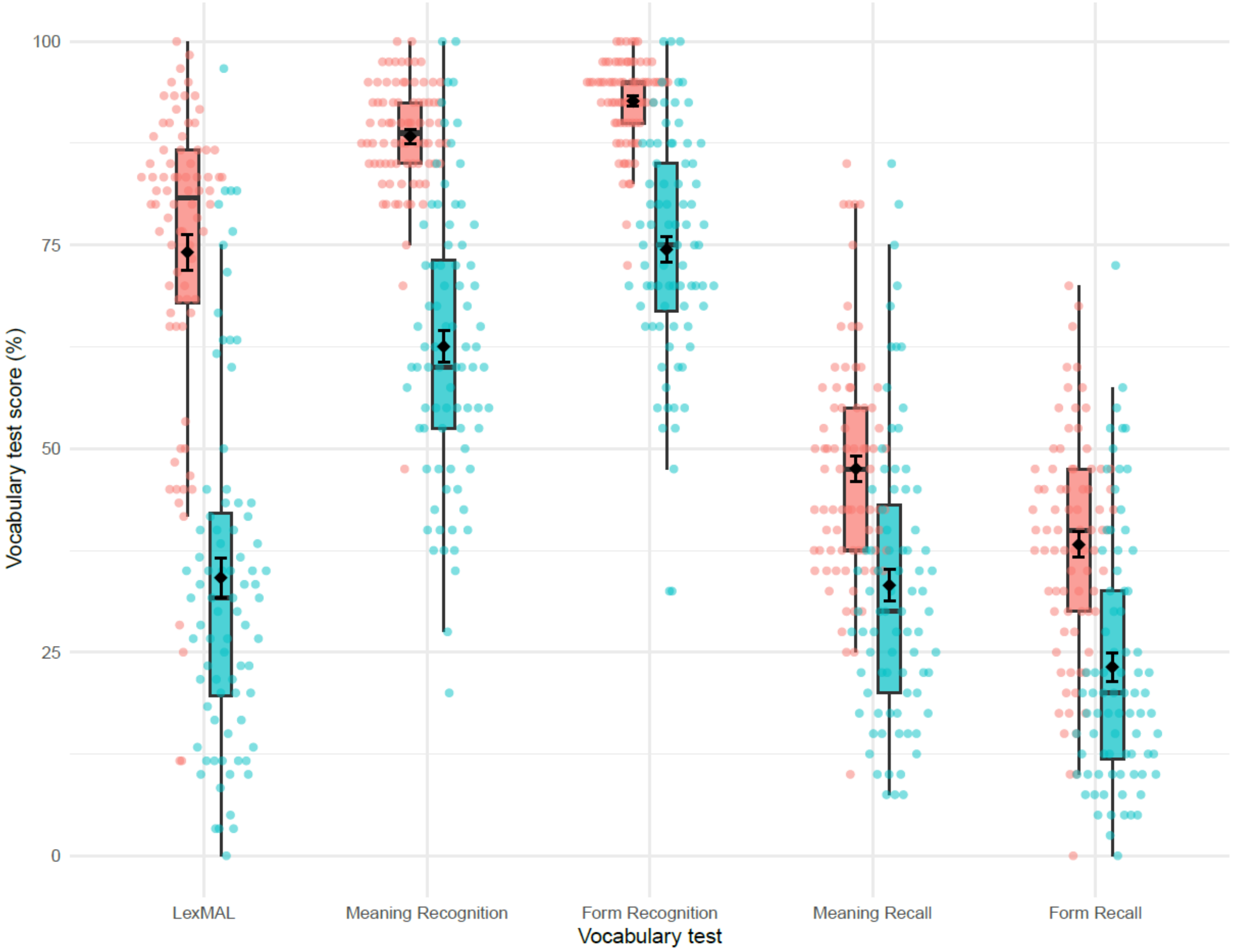
Note. Only one participant in the L1 group scored a perfect 100% in LexMAL.

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 4.1 for the by-participant plot). A fixed-effects hierarchical regression analysis was conducted to examine if the four vocabulary test scores predict LexMAL accuracy. Subsequently, a generalised 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 investigating language dominance effect across the vocabulary tests. The Malay L1/dominant speakers consistently outperformed the Malay L2/non-dominant speakers in LexMAL and other language tasks in Chapter 3. Therefore, this analysis included the categorical factor as control[14]. Lastly, the receiver operator characteristic (ROC) curve analyses (Lalkhen & McCluskey, 2008; Read et al., 2016) 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.

[14] We named the effect "language dominance" to appropriately reflect the linguistic background of our participants because the Malay L1 speakers in the present study not only acquired Malay as their first language, but also recognised it as their most proficient language and used it as their primary language in daily life.

Figure 4. 1

Vocabulary test scores across two language groups



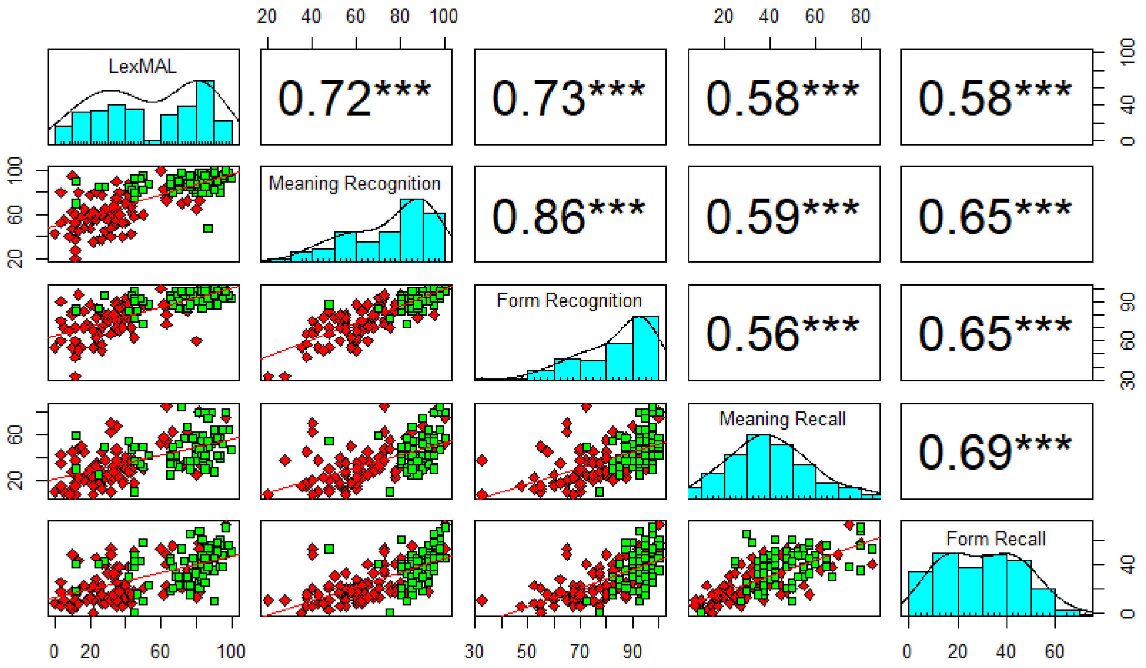
Note. The mean accuracy scores per participant group across vocabulary tests. Red represents the L1 speakers, and blue represents the L2 speakers. Black dots denote the group means, with standard errors denoted by the whiskers.

4.3.1 Predictive Power of Vocabulary Knowledge on LexMAL

Correlation analysis revealed that the scores of form-meaning vocabulary tests and LexMAL were positively correlated, *p*s < .001 (see Figure 4.2 for the correlation matrix). To assess if different mastery levels of form-meaning knowledge can account for a significant proportion of variance in LexMAL score, fixed-effects hierarchical regression analysis was conducted using R (version 4.1.1; R Core Team, 2021) with LexMAL score as the dependent variable and test scores from different mastery levels of form-meaning knowledge as fixed effects. The mastery levels of form-meaning knowledge were entered one-by-one into the model according to the acquisition order proposed in past studies, from the earliest to the latest (see Laufer & Aviad-Levitzky, 2017; Aviad-Levitzky et al., 2019; Laufer & Goldstein, 2004; Schmitt, 2010). Meaning Recognition score was entered in the first step to predict LexMAL score, followed by Form Recognition, Meaning Recall and Form Recall scores in the second, third, and fourth steps respectively.

Figure 4. 2

Correlation of scores between LexMAL and form-meaning vocabulary tests



Note. LexMAL scores were significantly correlated with all vocabulary test scores. Particularly, the correlations between LexMAL and recognition test scores were higher than those of recall tests.

The first three regression models explained significantly more variance than the previous models *F*s ≥ 11.89, *p*s < .001. The Form Recall score added to the final step did not account for additional variance in LexMAL score, *F* = 0.11, *p* = .74 (see Table 4.6 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. The variance inflation factors for Meaning Recognition, Form Recognition, and

Meaning Recall were 3.99, 3.79, and 1.56 respectively, suggesting that no collinearity issue was expected in the best fit model (variance inflation factors < 5).

Table 4. 6

Results of fixed-effects regression analysis

Variable	R^2 (adjusted R^2)	Estimate (*SE*)	*t* value
Step 1	0.53 (0.52)		
Meaning Recognition		1.12 (0.08)	13.22***
Step 2	0.57 (0.56)***		
Meaning Recognition		0.60 (0.16)	3.82***
Form Recognition		0.81 (0.21)	3.89***
Step 3	0.59 (0.59)***		
Meaning Recognition		0.47 (0.16)	2.98**
Form Recognition		0.72 (0.20)	3.55***
Meaning Recall		0.34 (0.10)	3.20**
Step 4	0.59 (0.58)		
Meaning Recognition		0.46 (0.16)	2.90**
Form Recognition		0.71 (0.21)	3.40***
Meaning Recall		0.32 (0.12)	2.63**
Form Recall		0.04 (0.14)	0.31

Note. *** $p < .001$, ** $p < .01$.

4.3.2 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 mastery levels (measured by form-meaning vocabulary tests) predict item accuracy, generalised mixed-effects modelling was conducted using the *lme4* R package (Bates et al., 2015). The fixed effects in the model were language dominance group (deviation coding[15]

[15] Deviation coding was employed to allow for inferential interpretations of main effects and main interactions. Deviation coding requires the values assigned to different levels of a factor to sum to zero. The target level is assigned the value of (k – 1)/k, where k represents the number of levels in a factor. On the other hand, the non-target levels are assigned the value of -(1/k). This coding scheme enables canonical ANOVA-style interpretations of main effects and interactions, which is in line with the research questions of the present study.

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. Because the scores from different vocabulary tests were moderately-to-highly correlated (rs ≥ .56; see Figure 4.2) and could induce collinearity concern, random intercepts and slopes were fitted with no correlation[16] (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 its interaction were included as by-item random slopes. The formula for the generalised mixed-effects model is provided below.

$$\begin{aligned} Item\ accuracy \sim Language\ dominance \\ &* (Meaning\ Recognition + Form\ Recognition + Meaning\ Recall \\ &+ Form\ Recall) + (1 + Meaning\ Recognition + Form\ Recognition \\ &+ Meaning\ Recall + Form\ Recall||participant) + (1 \\ &+ Language\ dominant * (Meaning\ Recognition + Form\ Recognition \\ &+ Meaning\ Recall + Form\ Recall)||stimuli) \end{aligned}$$

The generalised mixed-effects model revealed that language dominance affected vocabulary test accuracy (β = 1.70, SE = 0.18, z = 9.69, p < .001). For the same test items that

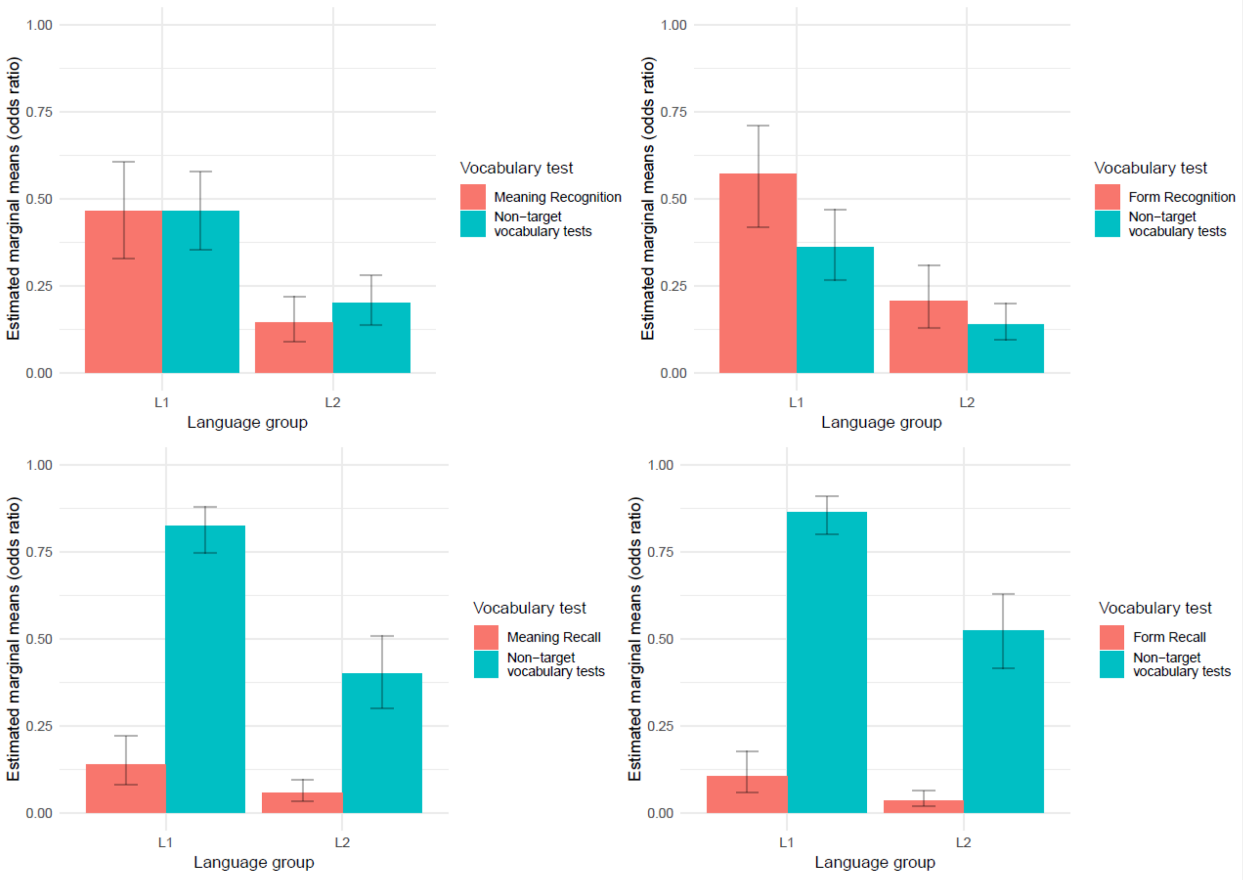
[16] This random effect structure helps to answer our research question if each of the vocabulary tests could predict item accuracy in LexMAL, instead of its unique contribution to predict LexMAL accuracy while taking other vocabulary tests into consideration.

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.

Main effect of vocabulary tests was 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 (β = 0.66, *SE* = 0.18, *z* = 3.58, *p* < .001), but lower in the Meaning Recall (β = -2.88, *SE* = 0.17, *z* = -17.40, *p* < .001) and Form Recall tests (β = -3.69, *SE* = 0.22, *z* = -16.47, *p* < .001). The log odds for Meaning Recognition was not significant (β = - 0.20, *SE* = 0.14, *z* = -1.43, *p* = .15), suggesting that there was no clear indication that correct identification of real words in LexMAL would predict their meaning being recognised in the Meaning Recognition test. Furthermore, significant interaction effects between language group and target test were found (see Figure 4.3) when the Meaning Recall (β = -0.99, *SE* = 0.23, *z* = - 4.37, *p* < .001) and Form Recall tests (β = -0.63, *SE* = 0.21, *z* = -2.96, *p* = .003) were examined as the target test against the other three non-target tests. The target test factor compares the odd ratios of the target and non-target vocabulary tests. The interactions between language group and the Meaning Recall as well as Form Recall tests were further examined in the next paragraph. Importantly, independent of the interactions, vocabulary items correctly identified in LexMAL were more likely to be correctly answered in Form Recognition test than other non-target vocabulary tests, regardless of language group. Table 4.7 provides an overview of the estimates of fixed effects and the interactions.

Figure 4. 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-right 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. 7

Summary of the generalised mixed-effects model

		Item accuracy	
Predictors	Odds Ratios	*CI*	*p*
(Intercept)	2.62	1.88 – 3.66	<**0.001**
Language group	5.47	3.88 – 7.71	<**0.001**
Meaning Recognition vs. LexMAL	0.82	0.62 – 1.08	0.153
Form Recognition vs. LexMAL	1.94	1.35 – 2.78	<**0.001**
Meaning Recall vs. LexMAL	0.06	0.04 – 0.08	<**0.001**
Form Recall vs. LexMAL	0.03	0.02 – 0.04	<**0.001**
Language group * Meaning Recognition	1.50	0.97 – 2.31	0.066
Language group * Form Recognition	1.47	0.91 – 2.36	0.114
Language group * Meaning Recall	0.37	0.24 – 0.58	<**0.001**
Language group * Form Recall	0.53	0.35 – 0.81	**0.003**
Random Effects			
σ^2	3.29		
τ_{00} participant	0.67		
τ_{00} stimuli	1.45		
τ_{11} participant.meaningrecognition	0.66		
τ_{11} participant.formrecognition	0.63		
τ_{11} participant.meaningrecall	0.63		
τ_{11} participant.formrecall	0.56		
τ_{11} stimuli.languagegroup	0.70		
τ_{11} stimuli.meaningrecognition	0.64		
τ_{11} stimuli.formrecognition	1.38		
τ_{11} stimuli.meaningrecall	1.20		
τ_{11} stimuli.formrecall	2.49		
τ_{11} stimuli.languagegroup:meaningrecognition	0.84		
τ_{11} 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.

Using *emmeans* R package (Lenth, 2023), post-hoc pairwise comparisons were conducted to examine how language group interacted with the target tests (i.e., Meaning Recall and Form Recall; see Table 4.8 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, *p*s < .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, *p*s < .001, indicating their poorer performance with the Meaning Recall and Form Recall tests. Specifically, the likelihood of scoring in the non-target tests was 29.35 times higher than in the Meaning Recall test and 54.57 times higher than in the Form Recall tests for the L1 speakers, indicating that the effects of target tests were stronger for L1 speakers than for L2 speakers (whose odds ratio was 29.18 at highest; see odds ratio in Table 4.8; implications of this finding are discussed in the Discussion section on page 184). The estimated marginal means and standard errors for each pairwise combination are summarised in Table 4.9.

Table 4. 8

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. *** *p* < .001, ** *p* < .01. T: target vocabulary test, NT: non-target vocabulary tests. Non-target vocabulary tests include all form-meaning vocabulary tests except the target vocabulary test.

Table 4. 9

Summary of estimated marginal means (EMM) and standard errors (SE) for each pairwise combination between language group, Meaning Recall and Form Recall

Pairwise comparison	Language group			
	Malay L1 (*N* = 80)		Malay L2 (*N* = 80)	
	EMM (odds ratio)	*SE*	*EMM (odds ratio)*	*SE*
Meaning Recall				
Target	0.14	0.04	0.06	0.02
Non-target	0.82	0.03	0.40	0.05
Form Recall				
Target	0.10	0.03	0.04	0.01
Non-target	0.86	0.03	0.52	0.06

Note. Non-target vocabulary tests include all form-meaning vocabulary tests except the target vocabulary test.

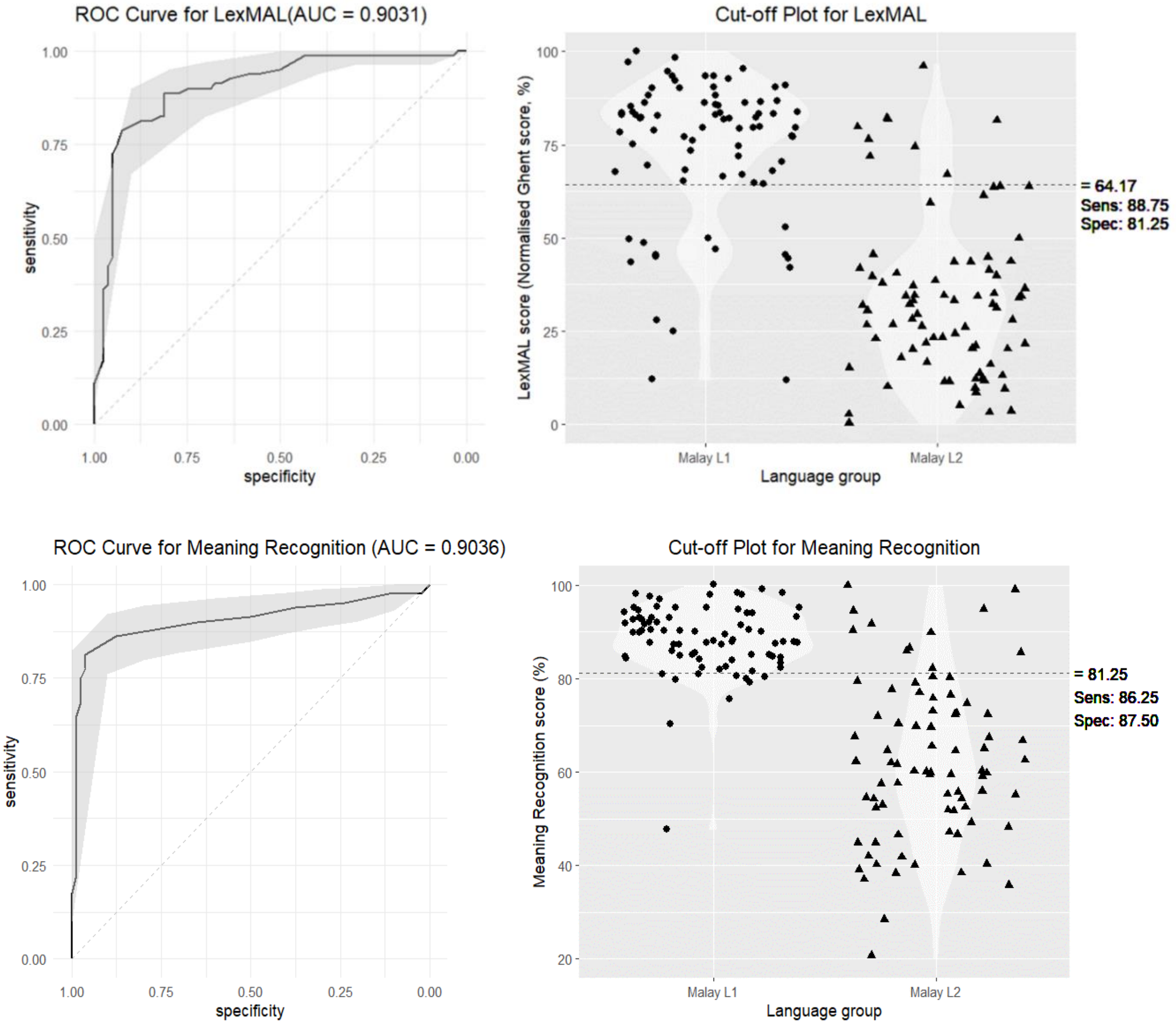
4.3.3 Discriminant ability and reliability of vocabulary tests

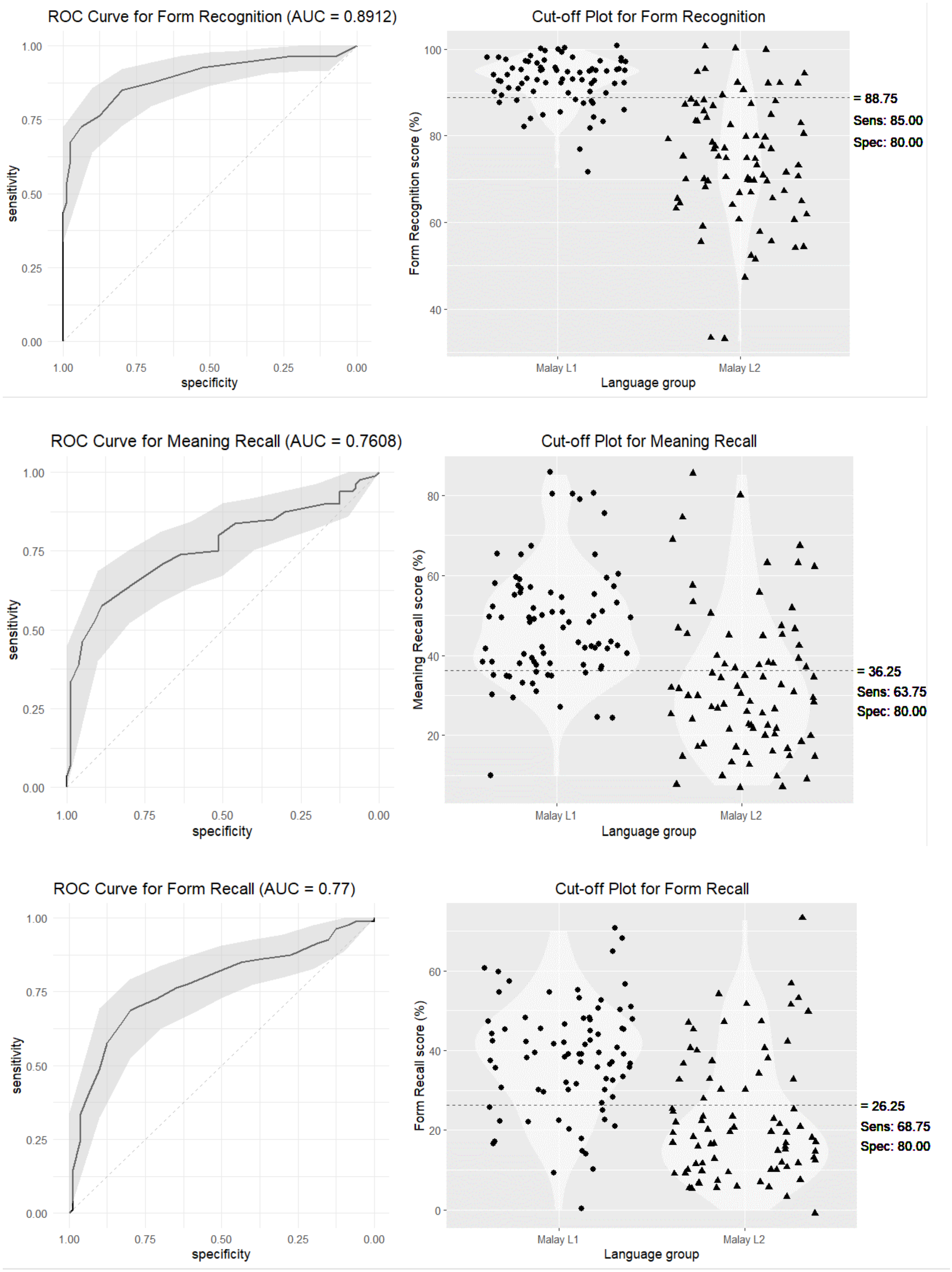
To examine if the vocabulary tests can distinguish L1 and L2 speakers' vocabulary knowledge, receiver operator characteristic (ROC) curve analyses were conducted using the *pROC* R package (Robin et al., 2021). ROC curve plotted the true positive rate (sensitivity) on the y-axis and the false positive rate (1 – specificity) on the x-axis. The vocabulary test's ability to discriminate between the vocabulary scores of L1 and L2 speakers is measured by area under the ROC curve (AUC). An AUC of .5 indicating no discrimination ability and an AUC of 1.0 indicating perfect discrimination (Hoo et al., 2017). In addition, the optimal cut-off score for each vocabulary test was also identified using point closest-to-(0, 1) corner method (see Chapter 3 for the same approach).

ROC curve analyses for the vocabulary tests (see Figure 4.4) revealed that LexMAL and Meaning Recognition test had very good ability in discriminating vocabulary knowledge of Malay L1 and L2 speakers, as indicated by their AUCs > .90. Form Recognition test's discriminant ability was also good with an AUC of .89. Meaning Recall and Form Recall tests, on the other hand, had fair discriminant ability with AUCs > .75. The optimal cut-off score for each vocabulary test was identified together with their sensitivity and specificity (see Figure 4.4 and Table 4.10). All vocabulary tests had Cronbach's alpha > .80, indicating good internal reliability (see Table 4.11).

Figure 4. 4

ROC curve for the vocabulary tests





ROC Curve for Form Recognition (AUC = 0.8912)

Cut-off Plot for Form Recognition

ROC Curve for Meaning Recall (AUC = 0.7608)

Cut-off Plot for Meaning Recall

ROC Curve for Form Recall (AUC = 0.77)

Cut-off Plot for Form Recall

Note. Sens: Sensitivity. Spec: Specificity.

Table 4. 10

Optimal cut-off score, sensitivity, and specificity of vocabulary tests

Vocabulary test	Cut-off score (%)	Sensitivity (%)	Specificity (%)
LexMAL	64.17	88.75	81.25
Form Recall	26.25	68.75	80.00
Meaning Recall	36.25	63.75	80.00
Form Recognition	88.75	85.00	80.00
Meaning Recognition	81.25	86.25	87.50

Table 4. 11

Internal reliability of the vocabulary tests

Vocabulary Test	*N*	Cronbach's alpha
LexMAL	90	.92
Meaning Recognition A	30	.90
Meaning Recognition B	30	.91
Form Recognition A	30	.87
Form Recognition B	30	.87
Meaning Recall A	30	.84
Meaning Recall B	30	.86
Form Recall A	30	.86
Form Recall B	30	.87

4.4 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/L2 speakers, AUCs > .75. Importantly, bilinguals' mastery levels of form-meaning knowledge, specifically form recognition, meaning recall, and form recall, were shown to predict bilinguals' item accuracy across the vocabulary tests.

At 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 score. 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 have its unique contribution in lexical proficiency (e.g., González-Fernández, 2022; González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; Nation, 2013, 2020, 2022). Meaning Recognition and Form Recognition tests, despite having high correlation between the test scores, measure distinct aspects of form-meaning knowledge under the unidimensional construct of vocabulary knowledge. Furthermore, meaning recall but not form recall explained a significant proportion of variance in the yes/no vocabulary test score. Because 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 prediction 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 imply that yes/no vocabulary test score could be a reliable indicator of bilinguals' receptive lexical proficiency because test takers' performance in the test corresponded well with their knowledge of meaning recognition, form recognition and meaning recall. This aligns with Bialystok's (2001) framework, emphasising that various linguistic tasks may impose different cognitive demands despite sharing a common underlying knowledge.

At item level, the generalised mixed-effects model revealed that different aspects of form-meaning knowledge were found to influence item accuracy on 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 in the latter tests, see Figure 4.3). The higher tendency for participants to recognise 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 correctly identified as real words, participants were more likely to be unable (than able) to recall their meanings or the word forms when their meanings were provided. This finding is surprising, as we did not expect negative predictions from any mastery levels of form-meaning knowledge. Nonetheless, as we will discuss in the following paragraphs, these negative predictions from form and meaning recall can be due to the higher difficulty level of the tasks compared to LexMAL. Evidently, across language groups, bilinguals' performances in the Meaning Recall and Form Recall tests were poorer than in the LexMAL, Meaning Recognition and Form Recognition tests. Taken together, it appears that there was a large knowledge gap between recognition and recall for the difficult words tested in the present study. Even when the bilinguals were able to recognise the word forms correctly, there is still a greater likelihood that they would not be able to recall the meanings or word forms.

Furthermore, correct identification of words in LexMAL does not indicate that test takers would be able to recognise 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 to participants' form recognition knowledge (Elgort, 2013; McLean et al., 2020).

The generalised mixed-effects model also revealed a significant difference in form-meaning knowledge between the two language groups, whereby the Malay L1 speakers outperformed the L2 speakers across all vocabulary tests. This is in-line with previous studies that reported L1 speakers to have larger vocabulary size than L2 speakers (Bialystok et al., 2008; Fernandes et al., 2007; Rahman et al., 2018). The L1-L2 speaker difference has also been consistently demonstrated in Chapter 3 and previous yes/no vocabulary test validation studies (Amenta et al., 2020; Brysbaert, 2013; Izura et al., 2014; Salmela et al., 2021; Wen et al., 2023), lending evidence to support 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 4.5). This is however in-line with the findings from our pilot study. The reason for this is two-fold. In-line with 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 in order to capture the variation in vocabulary knowledge of highly proficient L1 speakers (M_{Zipf} = 3.56, SD_{Zipf} = 0.54; see Table 4.2 for the distribution of target words across frequency bands). In addition to having a good blend of lexical decision difficulty (M_{accuracy} = 48.41%, SD_{accuracy} = 26.30%; taken from M. J. Yap et al., 2010), 50 out of the 60 target words (83.33%) have less than 50% translation accuracy by L1 speakers (M_{accuracy} = 24.44%, SD_{accuracy} = 21.95%; taken from Chapter 2). Because these target words were carefully selected from criterion-referenced norms gathered from highly proficient L1 speakers (as suggested by Bialystok, 2001), these difficult words were likely to tap into the higher language cognition of L1 speakers (Hulstijn, 2015, 2019). Put differently, by achieving the intended difficulty level of the tests, selecting word stimuli from criterion-referenced norms improves the validity of the test (Bialystok, 2001).

Because 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. Recognising the form and/or meaning of these words were easier for the L1 speakers when they were prompted by cues (e.g., recognising 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 were more difficult when they appeared in isolation or in a clueless context, even for the highly proficient L1 speakers. Compared to the L2 speakers, the L1 speakers had a larger knowledge gap between form-meaning recognition and recall (see Table 4.8). Despite their ability to recognise and recall more word forms and meanings than the L2 speakers, the L1 speakers' tendency to recall the word forms and meanings of the items they identified correctly in LexMAL did not catch up with their L1 advantage of having high accuracy in LexMAL and other recognition tasks. The recall tasks were apparently difficult even for the L1 speakers. In addition, this finding also 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 4.4), further indicating the importance of measuring L1 lexical proficiency in research (as demonstrated in Chapter 2; in accordance with Brysbaert et al., 2016; Hulstijn, 2015). Vocabulary tests like LexMAL (Chapter 3) and LexCHI (Wen et al., 2023), for example, could serve as a good tool for language research to measure L1 and L2 proficiency 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., at least 80% accurate) in identifying L1 (sensitivity) and L2 speakers (specificity). 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 cut-off 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 < .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 taken from Chapter 2; 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 cut-off scores. Taken together, our findings suggest yes/no vocabulary test and recognition tests to be better options than recall tests when the testing purpose is to distinguish the form-meaning vocabulary knowledge of L1 and L2 speakers or to identify speakers from a specific speaker group.

While all the vocabulary tests in the present study displayed high reliability and good discrimination ability, suitability of the test for research use depends on the purpose of 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. However, if the 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. Recognition tests are easier than the recall tests, therefore poses lower task demands on the participants. Specifically, the yes/no vocabulary test offers a quick and valid measure for lexical proficiency, whereby more items can be tested within a shorter period of time (c.f. Meaning Recognition and Form Recognition tests; see Table 4.4 for the summary of test duration). Given that the test scores from yes/no vocabulary test were positively predicted by form recognition but not meaning recognition knowledge, the test scores from yes/no vocabulary test to some extent capture test takers' ability to recognise some real word forms, even though they may not recognise the word meanings.

4.5 Conclusion

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 recognise the word forms given their meanings. Participants may not however, 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 score, 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.

Chapter 5

General Discussion

This thesis presented three empirical studies that focused on Malay-English translation in Malay-English bilinguals, measuring Malay language proficiency, and investigating the knowledge test takers used in yes/no vocabulary tests. The present chapter summarises the findings and focuses on the theoretical, methodological, and empirical contributions of the thesis to the literature and future research practice. Finally, this chapter will discuss the limitations of the research and provide recommendations for future research.

5.1 Summary of Main Findings

Chapter 2 presents the first empirical investigation into Malay-English bidirectional translation ambiguity by gathering translation equivalents from proficient Malay-English bilinguals. The prevalence of translation ambiguity between Malay and English is higher than that of other language pairs (e.g., Tokowicz et al., 2002; Prior et al., 2007; Tseng et al., 2014; Wen & van Heuven, 2017a). Importantly, the study expanded the scope of previous translation norming studies to also investigate a wider range of word classes, including nouns, verbs, adjectives, and word-class ambiguous items. Particularly, verbs were found to be more translation ambiguous than nouns in both translation directions. In addition, adjectives and word-class ambiguous items were at least as translation ambiguous as verbs. A consistent effect of within-language semantic variability on translation ambiguity is found in both translation directions (in-line with Allen & Conklin, 2014; Degani et al., 2016). In contrast, the relationships between lexical characteristics (i.e., word frequency and length) and translation ambiguity were inconsistent depending on language-specific properties of the language pair in question. In forward translation, bilinguals with higher L1 proficiency are more likely to provide correct and dominant translations in forward translation. L2 proficiency, however, was not correlated with either forward or backward translation accuracy. Overall, the Malay-English bidirectional translation norms allow for evidence-based decision making when choosing translation stimuli. The higher prevalence of translation ambiguity, inconsistent lexical characteristic and language proficiency effects when compared to previous translation norming studies highlight the importance of standardising the stimuli set to help investigating language-specific and language-universal factors affecting translation performance. Additionally, language proficiency should be measured across studies, ideally by a standard test to allow comparison across different participant groups.

Chapter 3 presents two experiments to address the predominant issue of the lack of language proficiency assessments for understudied languages. LexMAL, the first unspeeded yes/no vocabulary test in Malay, is developed to estimate Malay proficiency of L1 and L2 speakers. LexMAL was validated with four external criterion measures, including Malay-English bidirectional translation, cloze test, and self-rated proficiency. The validity of LexMAL was demonstrated through a reliable discrimination between L1 and L2 speakers, and significant correlations between LexMAL scores and performance on other Malay language tasks (i.e., translation accuracy and cloze test scores). In addition, LexMAL outperformed self-ratings in predicting cloze test accuracy. Reliability analysis found a high Cronbach's alpha of .94 (in-line with previous lextale-inspired studies, e.g., Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022). A validation study (Experiment 2) with the 90-item final LexMAL tested with a different group of Malay L1 and L2 speakers replicated the findings of Experiment 1. LexMAL takes less time to administer and score than translation tasks, and is less prone to ceiling performance than cloze tests. For research that involves Malay speaking bilinguals, LexMAL is a useful tool as a standard Malay proficiency measure, so that language proficiency of the bilingual participants can be accounted in the research findings, and allows for comparisons across different studies. LexMAL is freely available for researchers at www.lexmal.org.

Chapter 4 investigated the contribution of bilinguals' vocabulary knowledge, specifically knowledge of form and meaning, in yes/no vocabulary tests. Bilinguals' lexical proficiency was measured using LexMAL and four newly developed vocabulary tests. The findings revealed that 59% of the variance in the yes/no vocabulary test score was explained by the accuracy of Meaning Recognition, Form Recognition, and Meaning Recall tests, with knowledge of form recognition being the best predictor. Importantly, the item accuracy of yes/no vocabulary tests corresponds well to bilinguals' form recognition knowledge. Participants who correctly identified the word items in LexMAL also had a higher tendency to respond correctly in the Form Recognition test but not in the other form-meaning tests. The study provides evidence to support the use of yes/no vocabulary tests as a reliable lexical proficiency measure to estimate bilinguals' receptive language proficiency.

These findings collectively foster the understanding of language-specific and language universal processing by enabling cross-linguistic research in Malay-English language pair. The empirical studies also address existing research limitations in understudied languages by presenting an example of creating language proficiency tests in other languages. By assessing the vocabulary knowledge of L1 and L2 bilinguals who are highly proficient in both languages they speak, these findings also contribute to a better understanding of vocabulary knowledge in bilinguals. Moreover, the findings provide a template for vocabulary testing and demonstrates the utility of various vocabulary test formats in measuring L1 and L2 lexical proficiency. The following sections discuss each of these unique contributions.

5.2 Cross-Linguistic Research with Malay-English Language Pair

As a global lingua franca, English is extensively researched due to its high accessibility around the world. However, for a more comprehensive and inclusive understanding of universal language processing, greater linguistic diversity in language research is needed to prevent biases resulting from exclusive focus on English. For instance, lexical databases in non-English languages that provide linguistic information for cross-linguistic stimulus selection can be a good starting point (e.g., Peti-Stantić et al., 2021). This thesis contributes to the theoretical understanding about language-specific and language-general processing by studying the Malay-English language pair. As a member of the Austronesian language family, Malay offers a distinct contrast to English in cross-linguistic research, despite sharing the same 26 alphabetical letters (see Section 1.4.1 in Chapter 1).

In the past, research in non-English languages has been restricted by a lack of research tools and corpora available in the languages. For instance, there was no Malay-English translation norms available to guide stimuli selection for cross-linguistic experiments. Identification of appropriate translation equivalents for Malay-English cross-linguistic experiments was therefore challenging due to translation ambiguity issue (e.g., Allen & Conklin, 2014; Prior et al., 2007; Tokowicz et al., 2002; Tseng et al., 2014; Wen & van Heuven, 2017a; see also Section 2.1.1). This could potentially impact the validity of findings from Malay-English cross-linguistic experiments because language processing (e.g., translation recognition) can be affected by the type of translation equivalents used (e.g., dominant or non-dominant translation, Schwieter & Prior, 2020). For instance, using translation difficulty data from the Malay-English bidirectional translation norms allowed us to choose stimuli that were sufficiently challenging to capture variations among the highly proficient L1 speakers (as suggested by Bialystok, 2001, and Hulstijn, 2015, 2019). Thus, the first database of Malay-English bidirectional translation norms, coupled with comprehensive lexical and semantic information for both source words and their translation equivalents, presents itself as a remedy for the dearth of research resources in the language pair. The investigation of factors associated with Malay-English translation ambiguity and its comparison with other language pairs has improved our understanding of language-specific and language-universal bilingual processing.

5.2.1 Sources of Translation Ambiguity

Previous translation norming studies have revealed a high prevalence of translation ambiguity when English is paired up with Germanic (Dutch: Tokowicz et al., 2002), Romance (Spanish: Prior et al., 2007), and Sino-Tibetan languages (Chinese: Tseng et al., 2014; Wen & van Heuven, 2017a). When English was paired up with Malay, the translation ambiguity index was found to be even higher than all the other language pairs (e.g., Tokowicz et al., 2002; Prior et al., 2007; Tseng et al., 2014; Wen & van Heuven, 2017a). The language-specific source of translation ambiguity can be inferred from these comparisons because Malay is from a more distant language family (c.f. Dutch and English; see Section 2.4.1 for a detailed discussion). For instance, the higher translation ambiguity between Malay and English could be attributed to the language-specific conceptual mapping differences from one language to another (Schwieter & Prior, 2020; Tseng et al., 2014). Moreover, the effects of lexical characteristics on translation ambiguity were also specific to the language pair in question.

In addition to language-specific processing, the investigation of Malay-English language pair has also improved our understanding of the universal translation production process. Despite the variations in the typology of the language pairs, translation ambiguity in backward translation was consistently found to be higher than in forward translation across translation norming studies (e.g., Tokowicz et al., 2002; Prior et al., 2007), and semantic variability of the source words affected translation ambiguity in both translation directions (e.g., Allen & Conklin, 2014; Degani et al., 2016). These findings altogether suggest that the translation process from L2 to L1 is generally more ambiguous than the translation process from L1 to L2, and word meaning is accessed in both translation directions. These findings are useful to the development of word translation models. However, extending the findings to current models is not as straightforward. The challenges are discussed in the following subsection.

5.2.2 Unspeeded Translation Performance in Speeded Word Translation Models

Word translation models (e.g., Revised Hierarchical Model of Translation Ambiguity, Eddington & Tokowicz, 2013, Kroll & Stewart, 1994; Multilink, Dijkstra et al., 2019) have been developed to predict and provide explanations for bilinguals' translation performance. However, these models focus on speeded translation performance, with response latency and speeded accuracy as the measuring outcomes. For instance, the Revised Hierarchical Model (Kroll & Stewart, 1994; see also Eddington & Tokowicz, 2013) predicts a slower translation latency from L1 to L2 because forward translation requires access to the meaning of the L1 source words before a translation is produced in the L2. Backward translation, however, is relatively faster because it accesses the translation equivalents in the L1 before the meaning in L1 is activated. As L2 proficiency increases, the access of L2 words to their meanings became stronger, leading to a more balanced translation performance.

While these predictions of word translation models have been widely used to explain speeded translation production, it is difficult to apply them in explaining the findings obtained in this thesis (see also discussion in Section 2.1.3.3). Because unspeeded translation tasks do not require bilinguals to provide translations under time pressure, many other factors could affect their translation productions (e.g., careful consideration of the meaning to be translated for words with multiple senses, Degani et al., 2016). Therefore, interpretations of the behavioural outcomes and assumptions of the underlying cognitive processes involved in unspeeded translation are likely to be different from that of speeded translation. Thus, the empirical data from the translation norming studies calls for future word processing models to also include explanation for unspeeded translation production, and explain how various factors (e.g., lexical characteristics, language proficiency, conceptual mappings between languages) play a role in bilingual translation production.

5.3 Test Development for Understudied Languages

Language proficiency, despite being an important moderating factor in psycholinguistics and cognitive research, is not always objectively tested and reported in research (see Chapter 1 for a review). The systematic review by Park and colleagues (2022) revealed that only approximately 43% of English second language acquisition studies used an objective proficiency test to assess language proficiency, even though it is the most commonly investigated language in second language acquisition research. On the other hand, language proficiency testing for non-English languages varied substantially according to the target language. For example, at least 50% of studies used an objective language proficiency test for relatively frequently studied languages such as German, Spanish, and Dutch. Contrastively, understudied languages such as Finnish, Turkish, and Swahili were rarely objectively measured in past second language research, which may pose implications for interpretation of research findings.

Standardised vocabulary tests that underwent rigorous test development and validation were created as a practical solution for researchers seeking a freely available and quick language proficiency test. This helped to improve language proficiency testing in understudied languages that lack of research resources (e.g., language proficiency test and lexical databases). For instance, in many understudied languages (e.g., Finnish, Mandarin), unspeeded yes/no vocabulary tests (e.g., Lexize: Salmela et al., 2021; LexCHI: Wen et al., 2023) that were developed following the same creation criteria as LexTALE (Lemhöfer & Broersma, 2012) have enabled quick assessment of vocabulary knowledge as a proxy of language proficiency. By adapting the standard test development procedures from previous studies (e.g., Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Lemhöfer & Broersma, 2012; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022), this thesis demonstrates a case of rigorous item assessment and test validation procedures in developing such as test for Malay, an understudied language in Southeast Asia. In other languages where there is no reliable assessments of language proficiency for research, the procedures described in this thesis may provide a practical solution to create vocabulary tests.

5.3.1 Item Selection and Assessment

In the efforts of creating a practical and reliable yes/no vocabulary test under the circumstances of limited resources, this thesis presents a set of procedures needed for purposeful item selection and evaluation. Previous studies (e.g., Lemhöfer & Broersma, 2012; Wen et al., 2023) started with a larger set of words from a wide frequency range in the test prototypes and pruned it down to a smaller set of items in the final versions of the tests. Pilot studies were conducted with the target language users to evaluate the prototype items based on their item difficulty and discrimination power. Lemhöfer and Broersma (2012) in their pioneering work selected the final set of items based on the proportion of correct scores (difficulty level) and item-total correlations (discrimination power). Later, the lextale-inspired tests (e.g., Brysbaert, 2013; Izura et al., 2014; Wen et al., 2023) evaluated the items using the item response theory analysis. The final set of items were then selected based on item difficulty and discrimination power generated by the two-parameter logistic model.

The development of LexMAL used the same set of test development procedure from previous studies (Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Lemhöfer & Broersma, 2012; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022). At the first stage of LexMAL construction, the selection of a set of suitable word and nonword candidates was made possible with the availability of lexicon corpus (i.e., Malay Lexicon Project, M. J. Yap et al., 2010; see also Maziyah Mohamed et al., 2023). Using word frequency and lexical decision accuracy information from the Malay Lexicon Project, high frequency words that were likely to be known by most Malay speakers, as well as low frequency words that were more likely to be recognised only by highly proficient Malay speakers were selected for LexMAL prototype. Given the lack of a database that provides reliable nonwords for Malay lexical studies, pronounceable nonwords were created based on careful considerations. A nonword generator (i.e., Pseudo, van Heuven, 2020) was used to generate legal nonwords that obey Malay phonotactic rules based on existing Malay bigram and trigram frequencies. During item assessment (see Section 3.2.2.1), only items with positive point-biserial correlations were considered for the final LexMAL (86 words and 88 nonwords). Item response theory analysis was conducted to evaluate and select the final items that can measure both L1 and L2 proficiency and, at the same time, discriminate between the two groups of speakers. Based on their difficulty level and discrimination power, the final items (60 words and 30 nonwords) were selected to spread across a wide difficulty range and have the highest discriminative power at each difficulty level.

Word frequency and lexical decision accuracy have been commonly used in previous studies as reference for item difficulty, however its utility has not been formally assessed (e.g., Amenta et al., 2020; Brysbaert, 2013; I. L. Chan & Chang, 2018; Izura et al., 2014; Lemhöfer & Broersma, 2012; Salmela et al., 2021; Wen et al., 2023; Zhou & Li, 2022). To examine the usefulness of the lexical characteristics we used (i.e., word frequency and lexical decision accuracy) in selecting test items for lexical tests, an additional exploratory correlational analysis was conducted. Interestingly, in the final LexMAL, word frequency and lexical decision accuracy showed a negative correlation with item difficulty, *p*s ≤ .005 (see Table 5.1 for the correlation strengths), but not with item discrimination. These findings suggest that word frequency and lexical decision accuracy could be used as effective item selection criteria to gauge the test's difficulty, while how well these items could discriminate L1 and L2 speakers could not be informed by these lexical characteristics. Nevertheless, researchers can select test items based on these lexical information to accommodate the necessary difficulty level of the test, guided by the goal of the test they wish to create. Lexical databases that investigate language performances of language users in specific contexts (e.g., lexical decision, translation production) can be ecologically useful as a supplementary guide (Bialystok, 2001; Hulstijn, 2015, 2019). Following that, item response theory analysis can be used to prune the items based on their discrimination power.

Table 5. 1

Correlations between lexical characteristics, item difficulty, and discrimination power

Variable	1	2	3
1. Item difficulty			
2. Item discrimination	.06		
3. Word frequency	- .36**	.24	
4. Lexical decision accuracy	- .54**	- .18	.42**

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

5.3.2 Test Validation

Apart from careful item selection and assessment, this thesis gathered validity evidence for the LexMAL with multiple external criterion measures to justify the proposed use of the test. For many understudied languages, language tests are commonly developed in the absence of other available standardised tests that measure the same language construct. As a result, most studies (e.g., Brysbaert, 2013; Izura et al., 2014; Zhou & Li, 2022) only validated their tests with self-rated proficiency. Nevertheless, a well-developed and reliable vocabulary test should be validated with multiple similar or related criterion measures, preferably with at least one objective measure of language proficiency, as demonstrated in the validation of LexTALE (Lemhöfer & Broersma, 2012) and LexCHI (Wen et al., 2023). For instance, these yes/no vocabulary tests used bidirectional translation tasks as an external criterion measure for vocabulary knowledge (Lemhöfer & Broersma, 2012; Wen et al., 2023). Previous research has also used cloze tests from academic examinations as an additional criterion (e.g., Wen et al., 2023) in situations where a freely available academic language proficiency test is unavailable (c.f. the Quick Placement Test used in Lemhöfer & Broersma, 2012).

In accordance with Wen and colleagues' (2023) validation procedure, the validity of the LexMAL in assessing vocabulary knowledge and discriminating between L1 and L2 speakers was evaluated by a series of analyses. Bidirectional translation tasks, a cloze test, and self-rated Malay proficiency were used to validate the LexMAL (see Section 3.2.1.2). The group difference between the LexMAL scores of L1 and L2 speakers serves as key evidence for test validity, supporting LexMAL's utility in discriminating between the two groups of Malay speakers. Furthermore, the positive correlations between LexMAL scores and other measures of vocabulary knowledge (i.e., accuracy of Malay-English translation, English-Malay translation, and Malay cloze test) provide validity evidence to verify the suitability of LexMAL scores as a reliable indicator of vocabulary knowledge. Importantly, yes/no vocabulary test is shown to tap into form recognition knowledge within the multifaceted vocabulary knowledge construct (see Section 4.3.2), thereby offering an evidence base for test score interpretation.

Taken together, for understudied languages lacking standardised measures for test validation, criterion measures can be drawn from test formats that purport to measure the same construct (e.g., cloze test that assesses vocabulary knowledge). The successful use of translation tasks in Chapter 3 and self-devised form-meaning vocabulary tests in Chapter 4 supports the use of test formats with adequate face validity for such purpose. Adopting the validation procedures from previous studies (Lemhöfer & Broersma, 2012; Wen et al., 2023), the thesis provides evidence to support the validity of LexMAL scores as a measure of Malay lexical proficiency. Test validation should, however, continue to evaluate the extent to which the existing evidence supports the intended interpretation of test scores considering the intended use. With the current validity evidence as support, LexMAL is now freely available for research use to investigate other language processing (e.g., language processing latency and accuracy) that involves Malay users. This would further support the utility of LexMAL in psycholinguistics and cognitive research.

5.4 Bilingual Research with Early Malay-English Speakers

This thesis also contributes to the understanding of language proficiency effects in early bilinguals who are highly proficient in both languages they speak. Malaysia offers a unique and linguistically rich landscape for bilingual research because many Malaysians are early bilinguals, who either grew up speaking two languages (e.g., Malay and English) from birth, or acquired L2 before school age (see Section 1.4 for a review). Importantly, both languages are commonly used as lingua franca among the citizens, especially in the urban areas. This bilingual population is relatively rarely represented in research compared to late bilinguals who acquired their L2 later in adulthood (e.g., González-Fernández, 2022; Wiener & Tokowicz, 2021; H. Zhang et al., 2020). According to the Basic and Higher Language Cognition theory (Hulstijn, 2015, 2019; see Section 1.1.2 for a review), early bilinguals are likely to achieve L1-like proficiency in both languages they speak. They are likely to have acquired basic language cognition (i.e., knowledge of common vocabulary and syntax), and only demonstrate variation in higher language cognition (i.e., knowledge of low-frequency lexical items or uncommon morpho-syntactic structures) depending on their language experience such as level of education and leisure-time activities. This sets Malaysian bilinguals apart from the commonly studied late bilinguals who started learning an L2 after adulthood (e.g., through migration to a country where the L2 is the primary language). According to Hulstijn's theory, without sufficient quality exposure to the target language, it is more difficult for the late bilinguals to achieve high proficiency in their L2. For these bilinguals, they could demonstrate variation even in their basic language cognition.

5.4.1 Variability of L1 Proficiency in Higher Language Cognition

This thesis tested the prediction from Hulstijn (2015, 2019) with the unique population of Malaysian early bilinguals in the Malay-English translation norming study. Evidence presented in Chapter 2 points to significant variation in L1 proficiency among Malaysian bilinguals (see Section 2.3.3.1). This is in-line with Hulstijn's prediction for bilinguals' language ability, individual differences among the L1 speakers are expected when the higher language cognition is assessed. In the Malay-English translation norming study, bilinguals with higher L1 proficiency were more likely to provide correct and dominant translations in forward translation but not in backward translation. This is the first translation norming study that demonstrates L1 proficiency effects in the translation performance of highly proficient bilinguals. However, in contrast to previous research (e.g., Prior et al., 2007; Wen & van Heuven, 2017a), L2 proficiency was not correlated with either forward or backward translation accuracy. These findings seem to suggest that the translation performance of early bilinguals (who acquired both languages before school age) is different from that of late bilinguals who were commonly represented in previous translation norming studies (e.g., Prior et al., 2007; Wen & van Heuven, 2017a).

5.4.2 Stimulus Difficulty in Measuring Variation in Language Performance

The contrastive language proficiency effects observed between the two types of bilinguals are likely due to the difficulty of the translation stimuli. Because Malaysian early bilinguals in this thesis are highly proficient in both languages they speak, variation in language performance can only be captured with items that required knowledge of low frequency words or rare sentence structures (i.e., higher language cognition). For instance, the higher proportion of low frequency source words used in the forward translation task in Chapter 2 (see Table 2.6) required access to higher language cognition. Therefore, the source words were able to capture the variation in linguistic ability among L1 speakers (Hulstijn, 2015, 2019). Conversely, the majority of the source words used in the backward translation were high frequency words produced by highly proficient L2 speakers (from forward translation; see Section 2.2.2 for item sampling). Most of the highly proficient Malay-English bilinguals from the same language population are likely to find these words familiar. Therefore, translating these would only require basic language cognition. Because the early bilinguals are highly proficient in their L2, and therefore have acquired basic language cognition (Hulstijn, 2015, 2019), variation in translating these items could not be attributed to their language ability as estimated by LexTALE (Lemhöfer & Broersma, 2012) and self-rated proficiency.

The variation in higher language cognition among L1 speakers was later observed again in Chapter 3. When the translation stimuli selected were extremely difficult (based on translation error rates from the Malay-English translation norms), the L1 speakers demonstrated greater variance in their translation accuracy compared to other receptive tasks due to their variation in higher language cognition (see Figure 3.2). In addition, the range of their translation scores overlapped more with the L2 speakers in both translation directions, suggesting that heterogeneity in higher language cognition is shared by both L1 and L2 speakers. Therefore, language proficiency tests for research should also measure L1 proficiency, and low frequency words that tap into higher language cognition are useful for that purpose.

5.5 Measuring Lexical Proficiency of L1 and L2 Speakers

To enable accurate test score interpretation, rigorous test development should also clearly define the linguistic aspect used to estimate the language proficiency measured (Bialystok, 2001; Park et al., 2022; Schmitt et al., 2020; Schoonen, 2011). Assessing vocabulary knowledge can be complicated because the construct is made up of various interrelated but distinct aspects of word knowledge (Durrant et al., 2022; González-Fernández & Schmitt, 2020; Schmitt, 2010, 2014; Webb, 2013; see Section 1.3 for a review). For instance, an ongoing debate in vocabulary testing research revolves around determining the aspect of vocabulary knowledge that best predicts reading proficiency. While some researchers advocate for meaning recall tests (e.g., McLean et al., 2020; Stewart et al., 2023), others contend that meaning recognition tests serve as superior predictors (e.g., Laufer & Aviad-Levitzky, 2017). On the other hand, some argued that these test formats should be treated as multiple imperfect measures and used collectively to estimate vocabulary knowledge as a latent variable (Cromheecke & Brysbaert, 2022), which inevitably comes with a testing cost and might pose practicality issues. In addition to developing and validating a lexical proficiency test for the Malay-speaking population, this thesis also contributes to the conceptual understanding and knowledge of vocabulary testing. By clearly defining the conceptualisation of vocabulary knowledge, various test formats that assess different aspects of form-meaning knowledge were evaluated. This enables the investigation of relationships among the different aspects of form-meaning knowledge, subsequently provides recommendations for a reliable and valid vocabulary test based on the purpose of language testing.

5.5.1 Improving Language Proficiency Testing of Malay Bilingual Research

This thesis addressed the lack of language proficiency tests for understudied languages by presenting a set of procedures to develop a quick and valid Malay language proficiency measure for research use. LexMAL takes less time to administer and score than translation tasks, and is less prone to ceiling performance than cloze tests. For research that involves Malay speaking bilinguals, LexMAL can be a useful tool as a standard Malay proficiency measure that enables language proficiency of the bilingual participants to be accounted and compared across language studies.

The accumulated evidence in this thesis indicates that the discrete and context-independent yes/no vocabulary test (e.g., LexMAL) provides better and more consistent predictions compared to more subjective self-rated proficiency (in-line with Khare et al., 2013; Lemhöfer & Broersma, 2012; Tomoschuk et al., 2019; Wen & van Heuven, 2017a). While correlations between LexMAL scores and other vocabulary test scores were consistently replicated in this thesis, the correlations between self-rated proficiency and bilinguals' language task performances, on the other hand, were not always consistent. For instance, although significant correlations were found between self-rated L1 proficiency and Malay-English translation accuracy in the translation norming study and Experiment 1 of LexMAL validation study, the correlation, however, was not replicated in Experiment 2 of LexMAL validation study. Hence, despite being easy to collect, the research works reported in this thesis challenged the utility of self-rated proficiency because it is not always accurate and reliable (Brysbaert, 2013;

M. Li & Zhang, 2021; Tomoschuk et al., 2019). A meta-analysis revealed that self-rated proficiency only correlated moderately (*r* = .47) with externally measured language performance (e.g., vocabulary test scores), and only 20.43% of the variance in the objective language performance is accounted by the ratings (M. Li & Zhang, 2021). Furthermore, the accuracy of self-ratings is subject to between- and within-population differences (Tomoschuk et al., 2019). For instance, bilinguals of different language combinations (e.g., Spanish-English, Chinese-English), language dominance (e.g., Spanish-dominant, English-dominant), or language background (e.g., heritage speakers or recently immigrated bilinguals) were found to vary in their accuracy of self-rated proficiency (Tomoschuk et al., 2019). Therefore, while self-rated proficiency can serve as a useful complementary measure of language proficiency (in view of its practicality and moderate correlation with language performance), its interpretation can be problematic especially when comparison is being made across different bilingual groups, or bilingual subgroups with the same language combination.

On the other hand, LexMAL as a vocabulary test designed to measure Malay proficiency of L1 and L2 speakers consistently correlates strongly with other receptive vocabulary test scores (*r*s ≥ .69) and weakly-moderately with productive vocabulary test scores (*r*s ≥ .28). Furthermore, LexMAL scores consistently discriminated L1 and L2 speakers with large effect sizes (*d*s ≥ 1.92) across studies. LexMAL is thus superior to self-rated proficiency for assessing construct-specific lexical proficiency of bilinguals.

5.5.2 Measuring Vocabulary Knowledge for L1 and L2 Lexical Proficiency

Multiple vocabulary test formats were employed to measure lexical proficiency of Malay L1 and L2 speakers in this thesis. In general, all vocabulary tests (i.e., yes/no vocabulary test, cloze test, and form-meaning tests) were shown to have good internal reliability (see Section 4.11). The scores of the vocabulary tests were consistently correlated (see Tables 3.5 and 3.8; see also Figure 4.2), suggesting they tap into the same vocabulary construct when sum scores are considered.

5.5.2.1 A Distinction between Receptive and Productive Vocabulary Tests

While the correlations between receptive and productive vocabulary tests were consistent for L1 and L2 speakers as a whole, the correlations between yes/no vocabulary test and productive vocabulary test scores (i.e., translation production, or form and meaning recall) were not always consistent for the L1 speakers. This could be due to the variations in language proficiency levels reflected in the aspects of form-meaning knowledge measured by the tests (Bialystok, 2001). Because knowledge of recall is more difficult and acquired later than knowledge of recognition (González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004), it is not surprising that recognition test scores correlate with each other but not with recall test scores. However, despite the differences in the aspects of form-meaning knowledge measured, the scores from recognition and recall tests were expected to correlate because they both fall under a unidimensional construct of vocabulary knowledge (González-Fernández & Schmitt, 2020). The inconsistent correlations between the scores of LexMAL and productive vocabulary tests could also be due to the difficult translation stimuli selected for the LexMAL study (see discussion under Section 5.4).

In terms of practicality, LexMAL, Meaning Recognition and Form Recognition tests are time-efficient, taking about five to six minutes to complete (see Table 4.4). In terms of discrimination ability, the receptive vocabulary tests (i.e., LexMAL, Meaning Recognition and

Form Recognition tests) are more accurate than the productive vocabulary tests (i.e., Meaning Recall and Form Recall tests) in differentiating form-meaning vocabulary knowledge of Malay L1 and L2 speakers. Despite being argued to be superior to recognition tests in predicting reading proficiency (e.g., McLean et al., 2020; Stewart et al., 2023), productive tests including bidirectional translation tasks, Meaning Recall, and Form Recall tests only had a fair ability to discriminate between L1 and L2 speakers, with larger performance overlaps between the two groups of speakers and AUC values less than .80 (see Table 4.10 and Figure 3.2). They were also found to be less accurate in identifying L1 speakers, with sensitivity below 70%. Nonetheless, Meaning Recall and Form Recall tests had a specificity of 80%, which was higher than that of bidirectional translation tasks, where specificity were below 75%. These tests are still useful in capturing the effects of language dominance in modulating bilinguals' lexical proficiency (see Section 4.3.2; Bialystok et al., 2008; Fernandes et al., 2007; Rahman et al., 2018).

5.5.2.2 The Importance of Test Conceptualisation

Taken together, these findings highlight the importance of test conceptualisation in designing a vocabulary test. Given that different test formats assess different aspects of vocabulary knowledge (González-Fernández & Schmitt, 2020; Laufer & Aviad-Levitzky, 2017; Laufer & Goldstein, 2004; Nation, 2013, 2022; Read, 2000; Schmitt, 2010) and have different language demands (Bialystok, 2001), test format selection should align with its intended purpose and take into account the target learners, the context of the testing, and the specific aspects and levels of language constructs being measured (Schmitt et al., 2020). This would ensure that the resulting scores could reflect more accurately the test takers' language skills and provide useful information to users. For example, for vocabulary tests that purport to accurately distinguish the lexical proficiency of L1 and L2 speakers, receptive vocabulary tests with higher AUC values, together with good sensitivity and specificity should be chosen. On the other hand, for tests that targets both L1 and L2 speakers, difficulty levels of test items should be appropriately adjusted to include both easy and difficult items, so that the test would be able to capture the variability in both the highly proficient and less proficient speakers. In academic contexts where students' word learning or attainment is of interest, the vocabulary tests that demand direct demonstration of form and meaning knowledge might be more useful options for language teachers.

5.6 Limitations and Future Directions

The empirical studies in this thesis have improved our understanding of measuring bilinguals' unspeeded language performance through translation production, yes/no vocabulary test, and form-meaning vocabulary tests. However, the implication of these findings to inform existing language processing theories is not straightforward because many bilingual language models were developed with a focus on processing latency and speeded accuracy (see Section 5.2.2 for a discussion). Future research should therefore expand the existing online translation/word recognition models (e.g., Multilink, Dijkstra et al., 2019) to account for unspeeded translation performance. Along this line, the Malay-English translation norms and LexMAL can serve as a foundation to explore the relationship between speeded and unspeeded language performance in bilinguals. For instance, large-scale translation recognition and production experiments can be conducted using the same set of stimuli from the translation norms. By gathering translation latency and accuracy from Malay-English bilinguals, the relationship between unspeeded and speeded translation performance can be investigated. Similarly, LexMAL can be readily used to investigate the relationship between vocabulary size and word recognition speed, as well as to investigate response time as a function of lexical proficiency (e.g., Harrington, 2018).

LexMAL is validated with Malay speakers in Malaysia who are currently studying in or have graduated from tertiary education. Because vocabulary size can vary substantially depending on biographical-environmental factors and amount as well as types of literacy experiences (Brysbaert et al., 2016; Hulstijn, 2015, 2019), researchers who use LexMAL should be cautious when interpreting the test scores for Malay speakers of different age groups (e.g., younger adults), educational backgrounds (e.g., secondary school graduates), and other language backgrounds. For instance, given that Malay is the official language of four Southeast Asian countries in the Malay Archipelago (i.e., Malaysia, Singapore, Brunei and Singapore; Nomoto et al., 2018), the utility of LexMAL to Malay-speaking populations in other countries needs to be further investigated due to potential lexical variations. LexMAL can be used in conjunction with standardised language history questionnaires (e.g., P. Li et al., 2020; Marian et al., 2007) to evaluate its utility for Malay speakers with various language backgrounds, taking into account their language experience (e.g., country of residence) and current language exposure (e.g., amount of language usage in daily life).

LexMAL is designed to provide estimates of lexical proficiency for linguistics and psycholinguistics research use. Given its nature as a discrete vocabulary test, its predictive validity in relation to specific language skills (e.g., speaking or listening) has yet to be explored. Hence, it should not be employed as a replacement for in-depth academic language proficiency assessment for high-stake academic placement decisions. For instance, further investigation can be conducted to examine the extent to which LexMAL scores are useful in predicting students' attainment in Malay language classrooms. Therefore, validity evidence of LexMAL can be further evaluated in various real-life applications to justify its suitability and provide evidence-based score interpretation across different settings.

5.7 Conclusion

The thesis contributes to a greater linguistic diversity in linguistics and psycholinguistics research by offering a contrastive account from the Malay-English language pair. It achieves this through a comprehensive exploration of language processing and proficiency in Malay-speaking bilinguals. The empirical studies also address the scarcity of psycholinguistic research resources for Malay by presenting the Malay-English bidirectional translation norms and LexMAL. In addition, these studies established a valuable framework for the enhancement of vocabulary assessment practices by presenting a standard set of procedures for future test development and validation. Moreover, current scholarship in bilingual proficiency testing is further expanded by refining the interpretation of test scores in the context of yes/no vocabulary tests, pertaining to bilinguals' vocabulary knowledge. These contributions collectively represent significant advancements for language research in Malaysia, and potentially, the Malay Archipelago, contributing to a better understanding of bilingual language processing in a so far largely unexplored language.

References

Amenta, S., Badan, L., & Brysbaert, M. (2020). LexITA: A Quick and Reliable Assessment Tool for Italian L2 Receptive Vocabulary Size. *Applied Linguistics*. https://doi.org/10.1093/applin/amaa020

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Anderson, R. C., & Freebody, P. (1983). Reading Comprehension and the Assessment and Acquisition of Word Knowledge. *Advances in Reading Language Research*, *2*, 231-56.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of Success in Native and Non-Native Listening Comprehension: An Individual Differences Approach. *Language Learning*, *62*(s2), 49–78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

Allen, D., & Conklin, K. (2014). Cross-linguistic similarity norms for Japanese–English translation equivalents. *Behavior Research Methods*, *46*(2), 540–563. https://doi.org/10.3758/s13428-013-0389-z

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Basnight-Brown, D. M., Kazanas, S. A., & Altarriba, J. (2020). Translation ambiguity in Mandarin-English bilinguals: Translation production differences in concrete, abstract, and emotion words. *Linguistic Approaches to Bilingualism*, *10*(4), 559-586. https://doi.org/10.1075/lab.17037.bas

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge: Cambridge University Press.

Bialystok, E., Craik, F. I. M., & Luk, G. (2008). Lexical access in bilinguals: Effects of vocabulary size and executive control. *Journal of Neurolinguistics*, *21*(6), 522–538. https://doi.org/10.1016/j.jneuroling.2007.07.001

Bracken, J., Degani, T., Eddington, C., & Tokowicz, N. (2017). Translation semantic variability: How semantic relatedness affects learning of translation-ambiguous words. *Bilingualism: Language and Cognition*, *20*(4), 783–794. https://doi.org/10.1017/S1366728916000274

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing*, *33*(3), 307–318. https://doi.org/10.1177/0265532215583066

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., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997. https://doi.org/10.3758/s13428-012-0190-4

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., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bultena, S., Dijkstra, T., & van Hell, J. G. (2013). Cognate and word class ambiguity effects in noun and verb processing. *Language and Cognitive Processes*, *28*(9), 1350–1377. https://doi.org/10.1080/01690965.2012.718353

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One*, *5*(6), e10729–e10729. https://doi.org/10.1371/journal.pone.0010729

Chan, I. L., & Chang, C. B. (2018). LEXTALE_CH: A quick, character-based 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

Chan, S. H., & Abdullah, A. N. (2015). Bilingualism in Malaysia: language education policy and local needs. *Pertanika Journal of Social Sciences and Humanities*, *23*(3), 55–70.

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157–187. https://doi.org/10.1177/026765839401000203

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The Problematic Concept of Native Speaker in Psycholinguistics: Replacing Vague and Harmful Terminology with Inclusive and Accurate Measures. *Frontiers in Psychology*, *12*, 715843–715843. https://doi.org/10.3389/fpsyg.2021.715843

Cheng, S. H., & Lai, C. (Eds.). (2019). *Kamus Perdana: Bahasa Melayu-Bahasa Cina-bahasa Inggeris* (Edisi Keempat). United Publishing House (M) Sdn. Bhd.

Chinese-Malay-English Dictionary (Revised Edition). (2019). United Publishing House (M) Sdn. Bhd.

Clark-Carter, D. (2019). *Quantitative psychological research: the complete student's companion* (Fourth edition.). Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Lawrence Erlbaum.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review, 22*(5), 1216-1234. https://doi.org/10.3758/s13423-015-0819-2

Cromheecke, O., & Brysbaert, M. (2022). A French C-test for language assessment. *L'Annee Psychologique, 122*(1), 3–28. https://doi.org/10.3917/anpsy1.221.0003

Dang, T. N. Y. (2020). Vietnamese non-English major EFL university students' receptive knowledge of the most frequent English words. *VNU Journal of Foreign Studies*, *36*(3). https://doi.org/10.25073/2525-2445/vnufs.4553

Dash, T., & Kar, B. (2020). Behavioural and ERP correlates of bilingual language control and general-purpose inhibitory control predicted by L1 and L2 proficiency. *Journal of Neurolinguistics, 56*, 100914. https://doi.org/10.1016/j.jneuroling.2020.100914

De Groot, A. M. B. (1992). Bilingual lexical representation: A closer look at conceptual representations. In R. Frost & L. Katz (eds.), *Orthography, phonology, morphology, and meaning*, pp. 389–412. Amsterdam: Elsevier.

Degani, T., Prior, A., Eddington, C. M., Arêas da Luz Fontes, A. B., & Tokowicz, N. (2016). Determinants of translation ambiguity. *Linguistic Approaches to Bilingualism*, *6*(3), 290–307. https://doi.org/10.1075/lab.14013.deg

Degani, T., & Tokowicz, N. (2013). Cross-language influences: translation status affects intraword sense relatedness. *Memory & Cognition*, *41*(7), 1046–1064. https://doi.org/10.3758/s13421-013-0322-9

Degani, T., Tseng, A. M., & Tokowicz, N. (2014). Together or apart: Learning of translation-ambiguous words. *Bilingualism: Language and Cognition*, *17*(4), 749–765. https://doi.org/10.1017/S1366728913000837

Department of Statistics Malaysia. (2021, September). *Status e-Census Banci 2020 Sehingga 29 September 2021*. https://www.mycensus.gov.my/images/gallery/pdf/hightlightsOfMyCensus2020/29_September_2021.pdf

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. https://doi.org/10.1080/17470218.2012.720994

Dijkstra, T., Wahl, A., Buytenhuijs, F., van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition, 22*(4), 657-679. https://doi.org/10.1017/S1366728918000287

Dujardin, E., Jobard, G., Vahine, T., & Mathey, S. (2022). Norms of vocabulary, reading, and spelling tests in French university students. *Behavior Research Methods, 54*(4), 1611-1625. https://doi.org/10.3758/s13428-021-01684-5

Duñabeitia, J., Perea, M., & Carreiras, M. (2010). Masked Translation Priming Effects with Highly Proficient Simultaneous Bilinguals. *Experimental Psychology, 57*(2), 98-107. https://doi.org/10.1027/1618-3169/a000013

Durrant, P., Siyanova-Chanturia, A., Kremmel, B., & Sonbul, S. (2022). *Research methods in vocabulary studies*. Netherlands: John Benjamins Publishing Company.

Duyck, W., & Brysbaert, M. (2004). Forward and Backward Number Translation Requires Conceptual Mediation in Both Balanced and Unbalanced Bilinguals. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(5), 889–906. https://doi.org/10.1037/0096-1523.30.5.889

Earles, J. L., & Kersten, A. W. (2017). Why Are Verbs So Hard to Remember? Effects of Semantic Context on Memory for Verbs and Nouns. *Cognitive Science*, *41*, 780–807. https://doi.org/10.1111/cogs.12374

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa

Eddington, C. M., & Tokowicz, N. (2013). Examining English–German translation ambiguity using primed translation recognition. *Bilingualism: Language and Cognition, 16*(2), 442-457. https://doi.org/10.1017/S1366728912000387

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

Erlam, R. (2006). Elicited Imitation as a Measure of L2 Implicit Knowledge: An Empirical Validation Study. *Applied Linguistics*, *27*(3), 464–491. https://doi.org/10.1093/applin/aml001

Fang, N., & Zhang, P. (2021). L1 congruency, word frequency, collocational frequency, L2 proficiency, and their combined effects on Chinese–English bilinguals' L2 collocational processing. *International Journal of Bilingualism*, *25*(5), 1429–1445. https://doi.org/10.1177/13670069211024747

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using GPower 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fernandes, M., Craik, F., Bialystok, E., & Kreuger, S. (2007). Effects of Bilingualism, Aging, and Semantic Relatedness on Memory Under Divided Attention. *Canadian Journal of Experimental Psychology, 61*(2), 128-141. https://doi.org/10.1037/cjep2007014

Ferré, P., & Brysbaert, M. (2017). Can Lextale-Esp discriminate between groups of highly proficient Catalan-Spanish bilinguals with different language dominances?. *Behavior Research Methods*, *49*(2), 717–723. https://doi.org/10.3758/s13428-016-0728-y

Fox, J., Cheng, L., & Zumbo, B. (2014). Do They Make a Difference? The Impact of English Language Programs on Second Language Students in Canadian Universities. *TESOL Quarterly, 48*(1), 57-85. https://doi.org/10.1002/tesq.103

Fromont, L. A., 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. https://doi.org/10.1016/j.bandl.2020.104770

García, J. R., & Cain, K. (2014). Decoding and Reading Comprehension: A Meta-Analysis to Identify Which Reader and Assessment Characteristics Influence the Strength of the Relationship in English. *Review of Educational Research*, *84*(1), 74–111. https://doi.org/10.3102/0034654313499616

Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31*(1), 16–28. https://doi.org/10.1177/0734282912451971

Gentner, D. (1981). Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, *4*, 161–178.

Ginther, A., & Yan, X. (2018). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing, 35*(2), 271-295. https://doi.org/10.1177/0265532217704010

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

Gullifer, J. W., & Titone, D. (2018). Compute language entropy with {languageEntropy}. Retrieved from https://github.com/jasongullifer/languageEntropy

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition, 23*(2), 283-294. https://doi.org/10.1017/S1366728919000026

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

Harrington, M. (2018). Measuring Lexical Facility: The Timed Yes/No Test. In *Lexical Facility* (pp. 95–119). London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-37262-8_5

Harsch, C., & Hartig, J. (2016). Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing, 33*(4), 1-21. https://doi.org/10.1177/0265532215594642

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, *34*(6), 357. https://doi.org/10.1136/emermed-2017-206735

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

Hulstijn, J. H. (2010). Measuring second language proficiency. In E. Blom & S. Unsworth (eds.), *Experimental methods in language acquisition research* (EMLAR) (pp. 185–199). Amsterdam: John Benjamins.

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*. Amsterdam: John Benjamins.

Hulstijn, J. H. (2019). An Individual-Differences Framework for Comparing Nonnative with Native Speakers: Perspectives From BLC Theory. *Language Learning*, *69*(S1), 157–183. https://doi.org/10.1111/lang.12317

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

Ihlenfeldt, S., & Rios, J. (2023). A meta-analysis on the predictive validity of English language proficiency assessments for college admissions. *Language Testing, 40*(2), 276-299. https://doi.org/10.1177/02655322221112364

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

Jalil, S. B., Rickard Liow, S. J., & Keng, T. S. (2011). Semantic assessment battery for Malay-speaking adults with aphasia. *Aphasiology*, *25*(4), 415–433. https://doi.org/10.1080/02687038.2010.489259

Janebi Enayat, M., & Amirian, S. M. R. (2020). The relationship between vocabulary size and depth for Iranian EFL learners at different language proficiency levels. *Iranian Journal of Language Teaching Research*, *8*(2), 97-114. https://doi.org/10.30466/ijltr.2020.120891

Jasmani, F. (Ed.). (2013). *Kamus Cina-Melayu Dewan*. Dewan Bahasa dan Pustaka, United Publishing House (M) Sdn. Bhd.

Jasmani, F. (Ed.). (2012). *Kamus Melayu-Inggeris Dewan*. Dewan Bahasa dan Pustaka.

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

Jin, L., Razak, R., Wright, J., & Song, J. (2013). Issues in developing grammatical assessment tools in Chinese and Malay for speech and language therapy. In H. Winskel & P. Padakannaya (Eds.), *South and Southeast Asian Psycholinguistics* (pp. 145-156). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139084642.016

Jouravlev, O., & Jared, D. (2020). Native language processing is influenced by L2-to-L1 translation ambiguity. *Language, Cognition and Neuroscience*, *35*(3), 310–329. https://doi.org/10.1080/23273798.2019.1652764

Kementerian Pendidikan Malaysia. (2013). *Pelan Pembangunan Pendidikan Malaysia 2013-2025 (Prasekolah hingga Lepas Pendidikan Menengah).* Putrajaya: Kementerian Pendidikan Malaysia.

Keuleers, E. (2011). *Vwr: Useful functions for visual word recognition research* (0.1).

Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon*, *6*, 34–52. https://doi.org/10.1075/ml.6.1.02keu

Khare, V., Verma, A., Kar, B., Srinivasan, N., & Brysbaert, M. (2013). Bilingualism and the increased attentional blink effect: evidence that the difference between bilinguals and monolinguals generalizes to different levels of second language proficiency. *Psychological Research*, *77*(6), 728–737. https://doi.org/10.1007/s00426-012-0466-4

Klein-Braley, C. (1985). A cloze-up on the C-test: A study in the construct validation of authentic tests. *Language Testing*, *2*(1), 76-104. https://doi.org/10.1177/026553228500200108

Kroll, J., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections Between Bilingual Memory Representations. *Journal of Memory and Language, 33*(2), 149-174. https://doi.org/10.1006/jmla.1994.1008

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Fonseca, S. M. D., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Lõo, K., Marelli, M., Nisbet, K., Papadopoulos, T. C., Protopapas, A., Savo, S., Shalom, D. E., Slioussar, N., Stein, R., Sui, L., Taboh, A., Tønnesen, V., & Usal, K. (2023). Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition, 45*(1), 3-37. https://doi.org/10.1017/S0272263121000954

Kutlu, E., Tiv, M., Wulff, S., & Titone, D. (2022). Does race impact speech perception? An account of accented speech in two different multilingual locales. *Cognitive Research: Principles and Implications, 7*(1), 7–7. https://doi.org/10.1186/s41235-022-00354-0

Lai, C. (Ed.). (2018). *Kamus Kembangan* (Edisi Kedua). United Publishing House (M) Sdn. Bhd.

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

Laxén, J., & Lavaur, J. (2010). The role of semantics in translation recognition: Effects of number of translations, dominance of translations and semantic relatedness of multiple translations. *Bilingualism: Language and Cognition, 13*(2), 157-183. https://doi.org/10.1017/S1366728909990472

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, L. W., & Low, H. M. (2014). Analysis of Malay word structure by pre-service special education teachers: foundation-level knowledge for remedial instruction. *Australian Journal of Learning Difficulties*, *19*(1), 33–46. https://doi.org/10.1080/19404158.2014.891531

Lee, Y., Jang, E., & Choi, W. (2018). L2-L1 Translation Priming Effects in a Lexical Decision Task: Evidence From Low Proficient Korean-English Bilinguals. *Frontiers in Psychology*, *9*, 267. https://www.frontiersin.org/article/10.3389/fpsyg.2018.00267

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*, 1-17. 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.* https://doi.org/10.3758/s13428-023-02202-5

Lee, S. T., van Heuven, W. J. B., Price, J. M., & Leong, C. X. R. (under review). Assessing Bilingual Language Proficiency with Yes/No Vocabulary Test: The Role of Form-Meaning Vocabulary Knowledge.

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

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938-944. https://doi.org/10.1017/S1366728918001153

Ling, G., Wolf, M. K., Cho, Y., & Wang, Y. (2014). English-as-a-second-language programs for matriculated students in the United States: An exploratory survey and some issues. *ETS Research Report Series*, *2014*(2), 1-19. https://doi.org/10.1002/ets2.12010

Luniewska, M., Wodniecka, Z., Miller, C. A., Smolik, F., Butcher, M., Chondrogianni, V., Hreich, E. K., Messarra, C., A. Razak, R., Treffers-Daller, J., Yap, N. T., Abboud, L., Talebi, A., Gureghian, M., Tuller, L., & Haman, E. (2019). Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian. *PloS One*, *14*(8), e0220611–e0220611. https://doi.org/10.1371/journal.pone.0220611

Luque, A., & Morgan-Short, K. (2021). The relationship between cognitive control and second language proficiency. *Journal of Neurolinguistics*, *57*, 100956. https://doi.org/10.1016/j.jneuroling.2020.100956

Mahmud, F. N., & Salehuddin, K. (2023). How Bilingual Are Malaysian Undergraduates? A Snapshot of the Different Bilingual Categories in Malaysia. *GEMA Online Journal of Language Studies*, *23*(2). http://doi.org/10.17576/gema-2023-2302-08

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

Marian, V., Blumenfeld, H., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940-967. https://doi.org/10.1044/1092-4388(2007/067)

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

Maziyah Mohamed, 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

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. (1992). *New approaches to testing vocabulary knowledge*. Unpublished manuscript. Center for Applied Language Studies.

Meara, P. (1996). *English Vocabulary Tests: 10 k. Unpublished manuscript.* Center for Applied Language Studies.

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

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

Miller, G., & Fellbaum, C. (1991). Semantic networks of English. *Cognition, 41*(1), 197-229. https://doi.org/10.1016/0010-0277(91)90036-4

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

Mostafa, N. A. (2016). Bilingualism and education: The Malaysian experience. *Proceedings of EEIC*, *1*(2), 10-16.

Nakata, T., Tamura, Y., & Aubrey, S. (2020). Examining the Validity of the LexTALE Test for Japanese College Students. *The Journal of AsiaTEFL*, *17*, 335–348. https://doi.org/10.18823/asiatefl.2020.17.2.2.335

Nakatsuhara, F. (2011). The relationship between test-takers' listening proficiency and their performance on the IELTS speaking test. *IELTS Research Reports*, *12*(2011), 1. https://search.informit.org/doi/10.3316/informit.160409693787441

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian modern language review*, *63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

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

Nation, I. S. P. (2012b). Vocabulary Size Test instructions and description. Retrieved from https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

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

Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge: 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., & 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*. Boston, MA: Heinle, Cengage Learning.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: the validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*(3), 359–370. https://doi.org/10.1111/j.2044-8279.1997.tb01250.x

Neumann, H., Padden, N., & McDonough, K. (2019). Beyond English language proficiency scores: Understanding the academic performance of international undergraduate students during the first year of study. *Higher Education Research and Development*, *38*(2), 324-338. https://doi.org/10.1080/07294360.2018.1522621

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research: LTR, 21*(3), 298-320. https://doi.org/10.1177/1362168816639619

Ning, R. (2021). How language proficiency influences stroop effect and reverse-stroop effect: A functional magnetic resonance imaging study. *Journal of Neurolinguistics*, *60*, 101027. https://doi.org/10.1016/j.jneuroling.2021.101027

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"* (pp. 36–43).

Ockey, G. J., & Gokturk, N. (2019). Standardized language proficiency tests in higher education. In Gao X. (Ed.), *Second handbook of English language teaching* (pp. 1–17). Springer. https://doi.org/10.1007/978-3-319-58542-0_25-1

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, *32*(1), 39–62. https://doi.org/10.1177/0265532214538014

Oxford English-English-Malay Dictionary (3rd Ed. Updated Ver.). (2018). Oxford Fajar Sdn. Bhd.

Paek, I., & Cole, K. (2019). *Using R for item response theory model applications*. Routledge.

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*(1), 9–29. https://doi.org/10.18806/tesl.v11i1.623

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

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peti-Stantić, A., Anđel, M., Gnjidić, V., Keresteš, G., Ljubešić, N., Masnikosa, I., Tonković, M., Tušek, J., Willer-Gold, J., & Stanojević, M. (2021). The Croatian psycholinguistic database: Estimates for 6000 nouns, verbs, adjectives and adverbs. *Behavior Research Methods, 53*(4), 1799-1816. https://doi.org/10.3758/s13428-020-01533-x

Prior, A., MacWhinney, B., & Kroll, J. F. (2007). Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, *39*(4), 1029–1038. https://doi.org/10.3758/BF03193001

Prior, A., Wintner, S., Macwhinney, B., & Lavie, A. (2011). Translation ambiguity in and out of context. *Applied Psycholinguistics*, *32*(1), 93–111. https://doi.org/10.1017/S0142716410000305

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 Undergraduates. *3L, 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

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

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. (1993). The development of a new measure of L2 vocabulary knowledge. *Language testing*, *10*(3), 355-371.

Read, J. A. S. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Erlbaum.

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

Read, J. P., Haas, A. L., Radomski, S., Wickham, R. E., & Borish, S. E. (2016). 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

Rebuschat, P., & Mackey, A. (2013). Prompted production. In C. Chappelle (Ed.), *The encyclopaedia of applied linguistics* Vol. 5 (pp. 75–84). Oxford: Wiley–Blackwell.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., Siegert, S., Doering, M., & Billings, Z. (2021). *pROC: Display and Analyze ROC Curves*.

Rodríguez-Aranda, C., & Jakobsen, M. (2011). Differential contribution of cognitive and psychomotor functions to the age-related slowing of speech production. *Journal of the International Neuropsychological Society*, *17*(5), 807–821. https://doi.org/10.1017/S1355617711000828

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.

Rusli, Y. A., & Montgomery, J. (2020). Sentence Comprehension and Working Memory in Malay Adults. *GEMA Online® Journal of Language Studies*, *20*(1). http://doi.org/10.17576/gema-2020-2001-02

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. E., 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

Sawaki, Y., & Nissan, S. (2009). Criterion-related validity of the TOEFL iBT listening section. *ETS Research Report Series*, *2009*(1), i-82. https://doi.org/10.1002/j.2333-8504.2009.tb02159.x

Schmid, M., & Yılmaz, G. (2018). Predictors of Language Dominance: An Integrated Analysis of First Language Attrition and Second Language Acquisition in Late Bilinguals. *Frontiers in Psychology, 9*, 1306. https://doi.org/10.3389/fpsyg.2018.01306

Schmitt, N. (2010). Issues of Vocabulary Acquisition and Use. In *Researching Vocabulary. A Vocabulary Research Manual* (pp. 47–116). Basingstoke: Palgrave Macmillan. https://doi.org/10.1057/9780230293977_2

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

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

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.

Schoonen, R. (2011). How language ability is assessed. In E. Hikel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. II; pp. 701–716). New York, NY: Routledge.

Schwieter, J. W., & Prior, A. (2020). Translation Ambiguity. In *Bilingual Lexical Ambiguity Resolution* (pp. 96–125). Cambridge University Press. https://doi.org/10.1017/9781316535967.006

Shi, L. (2011). How "Proficient" is proficient? Subjective proficiency as a predictor of bilingual listeners' recognition of English words. *American Journal of Audiology, 20*(1), 19-32. https://doi.org/10.1044/1059-0889(2011/10-0013)

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

Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524636

Sigurd, B., Eeg-Olofsson, M., & van Weijer, J. (2004). Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, *58*(1), 37–52. https://doi.org/10.1111/j.0039-3193.2004.00109.x

Spit, S., Geambașu, A., Renswoude, D., Blom, E., Fikkert, P., Hunnius, S., Junge, C., Verhagen, J., Visser, I., Wijnen, F., & Levelt, C. C. (2023). Robustness of the cognitive gains in 7-

month-old bilingual infants: A close multi-center replication of Kovács and Mehler (2009). *Developmental Science, 26*(6), E13377. https://doi.org/10.1111/desc.13377

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*, Advance online publication. https://doi.org/10.1177/02655322231162853

Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the Vocabulary Size Test to a criterion measure of meaning recall vocabulary knowledge. *System, 87*, 102161. https://doi.org/10.1016/j.system.2019.102161

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

Szabo, C., Stickler, U., & Adinolfi, L. (2021). Predicting the academic achievement of multilingual students of English through vocabulary testing. *International Journal of Bilingual Education and Bilingualism, 24*(10), 1531-1542. https://doi.org/10.1080/13670050.2020.1814196

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

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). Amsterdam: John Benjamins.

Tokowicz, N., Kroll, J. F., Groot, A. M. B. de, & Hell, J. G. van. (2002). Number-of-translation norms for Dutch-English translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 435–451. https://doi.org/10.3758/BF03195472

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition*, *22*(3), 516–536. https://doi.org/10.1017/S1366728918000421

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. (2016) Language dominance: the construct, its measurement and operationalization. In C. Silva-Corvalan, & J. Treffers-Daller (Eds.), *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 235–265). Cambridge University Press.

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

Tseng, A. M., Chang, L.-Y., & Tokowicz, N. (2014). Translation ambiguity between English and Mandarin Chinese: The roles of proficiency and word characteristics. In J. W. Schwieter & A. Ferreira (Eds.), *The development of translation competence: Theories and methodologies from psycholinguistics and cognitive science* (pp. 107–165). Cambridge: Cambridge Scholars Publishing.

van Hell, J. G., & de Groot, A. M. B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition*, *1*(3), 193–211. https://doi.org/10.1017/S1366728998000352

van Heuven, W. J. B. (2020). *Pseudo* (2.10).

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 (2006)*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2011). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews*, *35*(3), 407–426. https://doi.org/10.1016/j.neubiorev.2010.04.007

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing, 29*(4), 603-619. https://doi.org/10.1177/0265532212448619

Wagner, E. (2020). Duolingo English Test, Revised Version July 2019. *Language Assessment Quarterly, 17*(3), 300-315. https://doi.org/10.1080/15434303.2020.1771343

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods*, *45*(3), 880–895. https://doi.org/10.3758/s13428-012-0289-7

Webb, S. (2005). Receptive and Productive Vocabulary Learning: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition, 27*(1), 33-52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2013). Depth of vocabulary knowledge. *The Encyclopedia of Applied Linguistics*, 346–354. https://doi.org/10.1002/9781405198431.wbeal1325

Webb, S. (2021). A Different Perspective on The Limitations of Size and Levels Tests of Written Receptive Vocabulary Knowledge. *Studies in Second Language Acquisition, 43*(2), 454-461. https://doi.org/10.1017/S0272263121000449

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). 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*. https://doi.org/10.3758/s13428-023-02151-z

Wen, Y., & van Heuven, W. J. B. (2017a). Chinese translation norms for 1429 English words. *Behavior Research Methods*, *49*. https://doi.org/10.3758/s13428-016-0761-x

Wen, Y., & van Heuven, W. J. B. (2017b). 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

Wiener, S., & Tokowicz, N. (2021). Language proficiency is only part of the story: Lexical access in heritage and non-heritage bilinguals. *Second Language Research, 37*(4), 681-695. https://doi.org/10.1177/0267658319877666

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *21*(2), 396–399.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497-528. https://doi.org/10.1177/0265532215594643

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

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

Yap, M. J., Sibley, D. E., Balota, D. A., & Ratcliff, R. (2012). Individual Differences in Visual Word Recognition: Insights From the English Lexicon Project. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10.1037/a0024177

Yap, N. T., Razak, R. A., Haman, E., Łuniewska, M., & Treffers-Daller, J. (2017). Construction of the malay cross-linguistic lexical task: A preliminary report. *Language Studies Working Papers*, *8*, 47–61.

Yi, W., & DeKeyser, R. (2022). Incidental learning of semantically transparent and opaque Chinese compounds from reading: An eye-tracking approach. *System, 107*, 102825. https://doi.org/10.1016/j.system.2022.102825

Zell, E., & Krizan, Z. (2014). Do People Have Insight Into Their Abilities? A Metasynthesis. *Perspectives on Psychological Science, 9*(2), 111-125. https://doi.org/10.1177/1745691613518075

Zhang, H., Jiang, Y., & Yang, J. (2020). Investigating the Influence of Different L2 Proficiency Measures on Research Results. *SAGE Open*, *10*(2). https://doi.org/10.1177/2158244020920604

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

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

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