Developing a Japanese Vocabulary Levels Test for the Purposes of Extensive Reading

KIMBERLEY ROTHVILLE

University of Auckland

When assessing L2 English learner vocabulary to guide learners to an appropriate reading level for extensive reading, teachers and researchers typically focus on high frequency words determined by analyses of L1 English corpora. This does not yet appear to be the case for Japanese, with teachers and researchers either focussed on textbook vocabulary lists or Japanese proficiency test vocabulary lists to determine learner vocabulary knowledge for either extensive reading or assessment purposes. This may be due to the lack of vocabulary tests based on Japanese word frequency. To partially address this gap, this paper reports on the creation of a vocabulary levels test based on Matsushita's (2012a) General Learners' Vocabulary List. The first iteration of this new vocabulary levels test, developed in 2017, covered only the first 2000 words, and was found to assess too few lexemes to be suitable for L2 Japanese learners. The current version was therefore expanded to cover the 5000 most frequent Japanese words. Four test forms were created, which have been used with second- and third-year learners at a large New Zealand university. The test was found to be highly reliable (Kuder-Richardson 21 = .98), and arguments are presented here for its validity in the L2 Japanese extensive reading context and for assessment of JFL learners' vocabulary.

Rothville, K. (2024). Developing a Japanese Vocabulary Levels Test for the purposes of extensive reading. *Journal of Extensive Reading*, *11*(3).

Keywords: Extensive reading, JFL, Vocabulary

There has been a growing awareness of extensive reading (ER) and its potential benefits for L2 Japanese learners within the Japanese teaching and research community over the past decade. However, there are a number of areas in which Japanese ER practice and knowledge lags behind that of the more innovative field of L2 English teaching and research. One of those areas is the use of vocabulary testing in order to match learners with suitable texts to read for ER and for assessing any vocabulary knowledge gains from ER. There appears to have been a lack of interest among Japanese language educators and researchers for developing vocabulary tests based on word frequency thus far. This is perhaps due to the existence of a list of specified vocabulary for the pre-2010 version of the Japanese Language Proficiency Test (JLPT) and the importance attached to passing the test as an endorsement of learners' Japanese language competence. That is, instead of testing learners on high frequency Japanese words, Japanese ER research appears to rely either on the JLPT vocabulary list (published by the Japan Foundation in 1994) or textbook vocabulary lists in order to assess learners' vocabulary knowledge. This is problematic for a number of reasons, but most prominently because neither the textbooks' lists nor the JLPT's vocabulary list are a good match for the actual frequency of words in texts written for L1 Japanese speakers. Learners will wish to begin reading L1 materials at some point in their language learning and extensive reading journey, so it is important to be able to assess their vocabulary knowledge against the frequency of words in text written for L1 speakers of Japanese.

Even when ER research does not deal directly with the development of learner vocabulary knowledge, matching of learners with texts to read appears to be based solely on three considerations: the stated level of the graded readers, whether pronunciation guides are present for kanji, or researchers' instincts about the "right" level for their participants. This potentially truncates the value of ER practices for L2 Japanese learners to benefit learner reading attitudes and other affective factors, as learners may become stressed or demotivated if texts are too easy or too hard for them to read. Additionally, English language learners need to ensure they know at least 95% of running words in texts they read, though ideally 98% or more if incidental vocabulary learning is desired (Nation, 2013). For Japanese, the proportion is unknown for incidental learning. Matsushita (2014) claims that learners should know 93% of running words for adequate comprehension with some dictionary help, while Komori, Mikuni, and Kondo (2004, as cited in Tamaoka, 2014) give a figure of 96% for general reading. The fine margins suggest that a gut-instinct about whether a text is at the "right" or "easy" level is not likely to be accurate, and a vocabulary levels test (VLT) should be used to assess learner vocabulary knowledge in order to match them with texts within their comprehension level. As such, a VLT is important because levels tests assess learner mastery of vocabulary at higher-frequency word bands, rather than assessing overall learner vocabulary size (Stoeckel, McLean, & Nation, 2021).

This paper introduces some of the main considerations for developing a vocabulary test and the decisions made when creating the Japanese Vocabulary Levels Test (J-VLT), as well as arguments for its validity considering the context and purpose of use, and pedagogical implications arising from having such a tool available in this field for the first time. This paper is intended only to cover the main considerations briefly, not to present a comprehensive validation of the tests in their current form.

Literature Review

ER and Vocabulary Knowledge

Extensive reading supports learner language acquisition through exposure to large amounts of target language input and has demonstrated benefits to language acquisition and to affective factors related to language study and foreign language reading (Yamashita, 2013). However, for extensive reading to be effective in these ways, it is important that learners read at the right level. Day and Bamford (1998) argue that learners should read either at their comprehension level (i) or slightly below their level of comprehension (i minus 1) in order to benefit from extensive reading. Although they note that learners may at some point be able to read comfortably just above their comprehension level, they assert that learners reading too far above their reading comprehension level or moving up too quickly may have the effect of increasing anxiety and demotivation, an approach to reading they term the "macho maxim" of reading pain equals reading gain (Day & Bamford, 2002, p.137). "Reading pain" can contribute to the development of foreign language reading anxiety, which has been noted to be significant for L2 Japanese learners (Saito, Horwitz & Garza, 1999). To avoid negative outcomes resulting from struggling to read at too high a level, it is important to be able to match learners with texts at their comprehension level.

One way to assess, in a very basic sense, whether a text may be appropriate for a learner is to match its vocabulary coverage level with a learner's vocabulary knowledge. Generally, for ER for L2 English learners, a text would be considered to be a candidate for a reader to be able to comprehend if 95-98% of words are known by that reader. Nation (2013) states that at minimum learners should already know 95% of running words in a text for reading comprehension, and 98% known words if the reader is interested in learning words incidentally from the text. However, without an appropriate test to measure learners' vocabulary knowledge, it is impossible to be certain if they are reading texts in which 95-98% of words are already known.

For Japanese, there does not appear to be published research that examines the proportion of running words in a text that need to be known by L2 readers for the purposes of ER activities. For general reading, Matsushita (2014) found an s-shaped relationship between learner vocabulary size and reading comprehension, but this may not be the case for ER. It is not clear, for instance, whether L2 Japanese learners need to read texts where 98% of words are known for incidental vocabulary learning, or if for Japanese there is a lower percentage coverage that would still be effective for learners engaged in ER.

Furthermore, for Japanese, only a small amount of graded material for L2 learners has been produced, meaning that for ER it is impossible to avoid having to supplement a programme with texts written for L1 speakers. It is not simply a case of adding children's books to a library's collection, as their vocabulary level is often as high as that of texts written for L1 adults (Webb & Macalister, 2013; Rothville, 2022), even if, at least for English, there appears to be suitable materials available (McQuillan, 2016). For effective ER it is therefore critical that, in addition to understanding learner vocabulary knowledge, the vocabulary level of texts used for ER can be analysed. Such a tool does exist: the J-LEX lexical profiler (Suganaga & Matsushita, 2013), but it is unclear how much awareness there is of it in the field, as no Japanese ER research appears to mention using it.

Measures of L2 Japanese Learner Vocabulary Knowledge

While for L2 English learners there is a wealth of research that examines the relationship between learner knowledge of frequent words as determined by examinations of corpora, the vocabulary coverage of texts, and vocabulary gains from ER, the same cannot be said for Japanese. Instead, research has concentrated on learner knowledge of the pre-2010 JLPT vocabulary list (see, for example, Horiba, Matsumoto, & Suzuki, 2006; Miyaoka, Tamaoka, & Sakai, 2011). However, this list does not reflect the frequency of words in texts written for L1 Japanese speakers (T. Matsushita, personal communication, 28th March, 2021) making relying on it problematic for measuring general learner vocabulary knowledge because it cannot be used to gauge how prepared L2 learners are to read texts written for L1 Japanese speakers.

For Japanese ER, it seems to be the case that there is little awareness either of the value of using frequency-based vocabulary lists, or of the necessity of assessing learner vocabulary knowledge to match them with texts that may be at their comprehension level. For ER research, primary considerations in published research thus far appear to be how many kanji are used (Hitosugi & Day, 2004; Leung, 2002; Tabata-Sandom & Macalister, 2009) or learner knowledge of the JLPT vocabulary (Fukumoto, 2004). Other ER studies do not appear to mention the method for matching learners with texts at their comprehension level, beyond comments noting books should be "easy" (see, for example, Banno & Kuroe, 2016).

Fukumoto (2004) and Leung (2002) appear to be the only published literature exploring L2 Japanese vocabulary gain from ER. Fukumoto examined learner vocabulary gains from a semester of ER as measured by the pre-2010 JLPT, finding the ER group (N = 21) increased their vocabulary and kanji knowledge score by nearly twice as much as the intensive reading group (N = 16). However, it was not clear how Fukumoto determined that only 2-5% of words in the text were unknown to the participants. Leung used a popular Japanese textbook to construct her vocabulary list, but given she studied from the textbook to some extent, it is unclear how much of the reported gains in vocabulary came from ER.

Research examining L2 Japanese learner vocabulary knowledge of the most frequent Japanese words appears to be very limited. Matsushita (2014) introduced the results of a Vocabulary Size Test (VST) for Japanese. Matsushita developed a VST of the 15,000 most frequent Japanese words and examined the relationship between learner vocabulary knowledge and reading comprehension in 213 adult learners studying Japanese in New Zealand, Australia, and Japan. Overall, there was a large range in terms of learner vocabulary knowledge scores (from less than 20 to greater than 130 out of a total of 150 items), and it was not clear how long learners had been studying Japanese or what their general proficiency level was.

Although Matsushita's VST represents a valuable first step in this field, there are three reasons why it may not be suitable for the purpose of assessing learner vocabulary knowledge for extensive reading or increases in L2 Japanese learner vocabulary knowledge. First, when using this test to assess learner vocabulary levels for the purpose of ER, a vocabulary size test is designed to estimate a learner's overall vocabulary in the target language, rather than their mastery at specific word frequency bands (Stoekel, McLean, & Nation, 2021). This means that it may not give appropriate information to enable understanding of learner vocabulary knowledge for the purpose of being able to read at a particular level. This is particularly true when sampling at a rate of, on average, ten words per 1000-word band (Gyllstad, Vilkaite, & Schmitt, 2015). Recent research suggests that sampling rates of less than 30 words per thousand produce a wide confidence interval of estimates of learner vocabulary knowledge, thus may under- or over-estimate learner knowledge to a significant degree (Gyllstad, McLean, & Stewart, 2020; Gyllstad, Vilkaite, & Schmitt, 2015). The second issue, for both ER and for assessing learner vocabulary knowledge or gains more generally, is that Matsushita's test does not appear to sample each word band evenly, nor are items from the same 1000-word band always grouped together, suggesting it would be difficult for teachers, learners, or researchers to adapt the test should they wish to focus on a smaller set of word bands. A measure of 15,000 words is more suited to more advanced learners, rather than beginner and intermediate level learners. Because it covers a high number of words that lower proficiency learners would not yet know, it is unclear if there is value in testing such learners on low frequency items. Third, the monolingual item format requires substantial knowledge of Japanese vocabulary and grammar in addition to the target word, potentially introducing confounding variables to the assessment of learner vocabulary, whether for ER or for testing learner vocabulary knowledge generally. There is therefore a clear need to develop a vocabulary levels test that focusses on a smaller set of vocabulary and samples vocabulary at a sufficient rate to produce a more accurate estimate of learner vocabulary knowledge, which is necessary if the goal is to match learners with texts to read for the purpose of ER. A vocabulary levels test would also need to have small enough bands, or have adjustable bands that can better match the vocabulary coverage levels of Japanese language resources for beginner readers, such as the various graded reader series that have been developed. Such a test would be more practical for learners, teachers, and researchers because it can be adapted to suit the learning context and purpose for testing.

Creating and Validating Vocabulary Tests

There are number of important considerations when creating a new vocabulary test (Schmitt, Nation, & Kremmel, 2020). In addition, for validating a vocabulary test, the purpose of a test, the people who will take it, and the educational context in which it is intended to be used are three of the most significant considerations when developing new vocabulary tests (Schmitt, Nation, & Kremmel, 2020). Without specifying these facets, a test's validity cannot be assessed, that is, it cannot be "determine[d] whether the test achieves its purpose or not" (Schmitt, Nation, & Kremmel, 2020, p.111).

The frequency list used to create the vocabulary test should be representative of domain to be investigated. That is, if researchers wish to examine learner knowledge of general, everyday words, the test should not sample from a list of academic vocabulary. On top of this, there is the choice of word counting unit. For L2 English learner vocabulary assessment, there is currently a debate on the appropriate counting unit to be used when testing vocabulary, and what the use of particular units says about what learners can do with the vocabulary they know (summarised in Webb, 2021). However, the current form of the debate regarding word units does not apply to Japanese, as the vocabulary system in Japanese is quite different to English due to the agglutinative nature of the language.

For Japanese, two frequency lists have been published: Matsushita (2012a) and Tono, Makawa and Yamazaki (2013). The two groups use slightly different terms (lexeme and lemma), but it appears that both refer to essentially the same method of dividing agglutinated Japanese words into their smallest meaningful units. That is, the only two options to sample from to create a frequency-based test use the 短単位 (short unit, similar to a lemma in English) as their word counting unit, so the discussion at this point in time is greatly simplified.

The choice of item format plays an important role in determining if a vocabulary test can meet its stated purpose. That is, if the knowledge of vocabulary measured by the test matches the goal of using the test. There are two main item formats: selected response (such as multiple-choice or multiple-matching formats, also termed meaning-recognition) and constructed response where learners are generally asked to provide a translation into their L1 or definition of the target word (also termed meaning-recall). Determining which is appropriate to use depends not just on what kind of vocabulary knowledge or degree of or strength of knowledge is to be assessed, but also the teaching and learning context of use, and the kind of vocabulary knowledge learners are developing. Additionally, teachers and researchers need to decide whether they would use the tests as a one-off assessment, such as for suggesting beginning reading levels, or if it will be an on-going assessment of learners' vocabulary development, in which case an item format that can detect vocabulary growth in the context of use is more suitable.

The translation, or meaning-recall, item format is generally regarded as more difficult than multiple-choice for learners because it asks them to recall from memory the meaning of a target word, thus assessing stronger or more well-developed vocabulary knowledge (Laufer & Aviad-Levitsky, 2017). A number of researchers assert that this is the depth or strength of knowledge required for reading (see, for example, Stoeckel, McLean, & Nation, 2021), though others dispute this to a certain extent (see, for example, Webb, 2021).

Another common item format for vocabulary tests is a selected response format, which has been used for both the English VST (Nation & Beglar, 2007) and the English VLT (Nation, 1990) for many years. Recent research has argued that this test type has issues with inflated scores from guessing (Gyllstad, Vilkaite, & Schmitt, 2015), and that it does not test the kind of vocabulary knowledge needed for reading (Stoeckel, McLean, & Nation, 2021), although Laufer and Aviad-Levitsky (2017) suggested there was no difference between either meaning-recognition or meaning-recall tests to predict general reading ability. Other researchers have argued that the fact that multiple-choice formats reward partial vocabulary knowledge is not necessarily detrimental, as it will detect learners' developing vocabulary knowledge (Webb, 2021), which can make such tests useful for assessing learner vocabulary development. Although it may be the case that it is possible to guess correctly on multiple-choice tests, potentially causing estimates of learner vocabulary knowledge to be inflated, it does appear that this estimation error is unidirectional (Kremmel & Schmitt, 2016). That is, learner knowledge is only overestimated, rather than there being a combination of over- and under-estimating of vocabulary knowledge, as there is for translation tests, which makes it potentially easier to adjust for the error (Kremmel & Schmitt, 2016).

Furthermore, multiple-choice item formats have an advantage in that students know exactly what they need to do in order for their responses to be marked correct. This contrasts with translation (meaning-recall) formats, where learners may not know how much detail to give in their answers if they do not know an exact translation of the target word and thus may not receive credit for words they do know, as occurred in the experiments reported in Kremmel and Schmitt (2016). Another advantage of multiple-choice tests is that they are less challenging to mark than translation tests, as a certain amount of subjective judgement will be present when scoring marginally correct answers. For this reason, multiple-choice tests may be slightly more reliable (Webb, 2021). McLean et al. (2021) argue that automatically marked meaning-recall (translation) tests can achieve a higher internal consistency that those scored by human markers. Yet, their paper still demonstrated that there were cases where the program marking their test incorrectly graded responses as correct, and where it incorrectly graded responses as incorrect, thus necessitating the manual checking of answers to ensure the accuracy of the logged scores.

Finally, there is the number of items the test needs to include from each word band. For ER in English, learners should know 95% of words in each word band to read texts at that level, or 98% in order to learn new words from texts at that level, meaning that a test should include sufficient items in each band to be able to calculate learners' scores to that level of detail. Although it is not yet clear what level of vocabulary coverage is necessary for JFL learners when engaged in ER specifically, it has been suggested that a minimum of 93% of words in a text should be known by learners in order to read with some dictionary help (Matsushita, 2014), or 96% for general reading (Komori, Mikuni, & Kondo, 2004, as cited in Tamaoka, 2014). This indicates that based on the current understanding of the relationship between learner lexical knowledge and reading comprehension, a Japanese Vocabulary Levels Test should include sufficient items in each band so that learners' knowledge can be estimated to at least this level of detail. For tests that aim to measure learner vocabulary gains, gains from incidental and classroom learning are marginal, being in the range of four words for each classroom hour, or a few percent increase for incidental learning (Gyllstad, Vilkaite, & Schmitt, 2015; Stoeckel, McLean, & Nation, 2021). A test should include at least 30 items per 1000-word band in order to properly represent the underlying population of words and to capture such incremental gains (Gyllstad, McLean, & Stewart, 2020; Gyllstad, Vilkaite, & Schmitt, 2015).

Fundamentally, very little L2 Japanese vocabulary research exists, either relating to learner vocabulary knowledge or developing tests to assess various aspects of learners' knowledge. Thus, it has been necessary to draw to a certain extent on the existing literature for L2 English vocabulary research, with the caveat that the two languages have very different vocabulary systems. The first major difference is that it is generally straightforward to tell if a word is a noun, adjective, verb, or adverb in Japanese. Each category tends to have distinct and consistent morphological patterns (Kageyama & Kishimoto, 2016; Miyaji, 1969), which learners can readily identify. That is, when an English language learner encounters the word "run" for the first time, whether as part of a text or on its own, they will not be able to tell whether this is a verb, noun or adverb, for example. In contrast, when learners of Japanese encounter the same word 「走る」[run], even if they do not know the meaning, they can assume that this new word is a verb due to its ending. This means that learners of Japanese can almost always interpret the role or function of an unknown Japanese word without knowing its meaning, unlike learners of English.

The other major difference is the semantic cues found in the kanji characters, and these make up around 40% of Japanese texts, rising to two-thirds when only content word tokens are considered (Matsushita, 2014). Japanese kanji tend to be composed of two parts, one denoting the semantic domain of the kanji, and the other giving a clue to one of the ways it may be pronounced (Kess & Miyamoto, 1999), though it should be noted this pronunciation guide is not very reliable. Kanji with this semantic component make up more than 80% of Japanese kanji (Kess & Miyamoto, 1999). That is, when an English language learner encounters the unknown word "shark", for instance, they cannot tell even vaguely what the meaning might be or relate to, but this is not the case in Japanese. A learner of Japanese encountering 「鮫」[shark] would see the left side of the kanji in composed of 魚 [fish] and would be able to guess that this unknown word has something to do with fish, or that it may be some kind of fish. Although this is not foolproof (for example, 「鮮やか」[fresh]), it still provides useful cues as to the semantic domain of unknown Japanese words, thus aiding their comprehension by learners in a way that is fundamentally different to the situation English language learners may find themselves in when confronted with an unknown vocabulary item.

In sum, even beginner learners of Japanese may already be able to interpret the part of speech and therefore the function of unknown Japanese words, and for words containing kanji, they may be able to infer information about the semantic domain of that word. That is, a learner of Japanese encountering an unknown word either while engaged in a reading activity or during a vocabulary test has more information they can use to interpret potential meanings of an unknown word available to them compared to a learner encountering an unknown word in English.

Given these significant differences, it should be explicitly stated that even if a finding has been accepted to be generally true for L2 English vocabulary research, such findings are not necessarily transferrable to other languages such as Japanese. Considerable work is still needed in the emerging field of L2 Japanese vocabulary research, not only to develop tools for assessing L2 learner vocabulary, but also to understand which tools best measure learner vocabulary knowledge and development in different contexts and for different purposes. While L2 English vocabulary researchers may appear to agree that for general reading and extensive reading, translation tests perform better, the question has not been answered for L2 Japanese learners. At this point, neither translation (meaning-recall) tests, multiple-choice (meaning-recognition), nor an appropriate reading comprehension test exist or have been published. That is to say, neither translation nor multiple-choice formats should be discounted for use at this very early stage in the development of this field, as the tools to probe the question of which format performs better have only just begun to be developed. Testing students in large enough numbers to lay the empirical foundations to answer this question is still required.

To sum up, research exploring L2 Japanese ER has been hampered by the lack of a Japanese Vocabulary Levels Test (J-VLT) to assess whether learners know enough words to read material at a particular word level. A VLT differs from a VST, which estimates a learner's overall vocabulary size. A VLT enables the estimation of a learner's vocabulary within a particular word band, which are usually set at 1000 words for L2 English learners. In addition, a vocabulary test is needed that can measure the incremental gains from incidental learning as a result of ER to better understand how ER can support L2 learner language acquisition. The development of a VLT based on word frequency in L1 Japanese text is therefore of critical importance to enable the field of Japanese ER to progress in understanding the role ER plays in facilitating this aspect of language acquisition for Japanese, as well as for matching learners with texts that are at or below their comprehension level. This paper lays out the rationale and methodology for constructing the frequency-based VLT for Japanese, in terms of the current multiple-choice format of the test. A translation (meaning-recall) version of the J-VLT has been created and is undergoing further refinement.

Methodology

In order to address the lack of a Vocabulary Levels Test for L2 Japanese learners by developing one, a number of issues need to be considered. Firstly, there is the choice of frequency list used to create the items for the test, and how that relates to the potential material learners may be reading. Secondly, there is the choice of item format and whether the format used assesses the right kind of knowledge for the context the test will be used in. Third, there is the rate of sampling of vocabulary items to be tested, as well as how many word bands the list will cover. Finally, though especially important for Japanese vocabulary, there is the problem of creating suitable distractors for multiple-choice items.

In addition, it is important to begin with a brief overview of the purposes for which the test could be employed and the contexts in which it could be used. Firstly, the context of use for which the test is developed is an input-poor foreign language environment, that is, learning contexts outside of Japan where there is little exposure to the target language outside the language classroom. Even if students and teachers are interested in extensive reading in Japanese, there are few texts available for beginner readers to read and thus the total input available may not be sufficient to enable learners to develop a high strength or depth of vocabulary knowledge (Abe, 2016; Rothville, 2022). These context constraints have led to the choice of a meaning-recognition item format to measure learner vocabulary knowledge and growth. This means that the test in its current form is not designed to measure learners' vocabulary knowledge for productive language use (e.g., speaking or writing).

Secondly, the test is designed to be used with adult, tertiary-level learners, although it may be possible that it is suitable for older teenage learners or those self-studying Japanese. Its suitability for use in pre-tertiary contexts will likely depend on learners' English vocabulary knowledge as there are a number of words in the test that L1 English speaking children may not be familiar with.

Thirdly, there are two purposes for which it might be used. The first purpose is to assess learner mastery of word bands at the meaning-recognition level of knowledge in order to decide if texts at a particular level may be within their ability to read extensively, rather than to study intensively. Until now, L2 Japanese learners have not had any practical means by which to estimate their vocabulary knowledge relative to the most frequent Japanese words, and thus to match themselves with Japanese texts at the same level. The second purpose is to estimate learner vocabulary growth, given that learners may not be encountering vocabulary enough times either in the foreign language classroom environment or while reading extensively to develop sight vocabulary and an instrument that can assess learners' growing familiarity and reward partial knowledge will be more suitable in the foreign language learning context. While the above cited literature for L2 English learner vocabulary knowledge may be settling on the position that meaning-recall (that is, translation) formats are better predictors for learners' ability to read in English (importantly, this relates to reading in high-stakes exams, not ER), there is not yet evidence for L2 Japanese learners as to whether knowledge demonstrated through translation or through multiple-choice tests is the better fit for the purpose of ER in Japanese or learner vocabulary gains in the foreign language context. This is not to say that the intent here is to argue for one format or the other, but rather to assert that the development of this test represents a first step by which researchers in this field can begin to probe such questions.

Choice of Frequency List Used as Base Population for the Test Sample

The test here is constructed from the General Learner's List, which lists the first 20,000 most frequent words in order of frequency and was constructed from the Balanced Corpus of Contemporary Japanese by Matsushita and published in 2012 as supplementary data alongside his PhD thesis. Matsushita made a pedagogically-motivated decision to move the 1288 words that form the vocabulary list of the lowest two levels of the pre-2010 JLPT to the beginning of his frequency list, as he determined that they were still useful for beginner learners to some extent (T. Matsushita, personal communication, 28th March, 2021). Although the first 1288 words are out of order of frequency, it was considered a better option than the list published by Tono, Makawa and Yamazaki (2013) in their Frequency Dictionary of Japanese: Core Vocabulary for Learners because of the existence of the J-LEX Japanese lexical analyser website published by Suganaga and Matsushita (2013), which allows users to check the vocabulary coverage of Japanese texts according to Matsushita's vocabulary list. This means that by using Matsushita's list, learners can be matched with texts based on their vocabulary knowledge, rather than using guesswork, which is important for the purposes of ER.

In some ways this decision regarding the abovementioned 1288 words may be problematic for learners, as the JLPT list does not match the frequency of words used in L1 Japanese texts. However, because being able to match learners to texts at the 95%, 98%, or 100% vocabulary coverage level is regarded to be essential for the purposes of ER, incidental vocabulary acquisition, or reading fluency development, the ability to pair the tests with output from the lexical analyser (J-LEX) was a paramount consideration. On the other hand, a significant proportion of the currently published graded readers appear to use the JLPT list as the basis for their vocabulary, suggesting it may not be as detrimental as it first seems, as these words are likely to be used with high frequency in the lower-level graded readers that learners read when beginning ER for the first time.

Sampling

The Japanese Vocabulary Levels Test was first developed in 2017 for research reported in Rothville (2019) which explored the impact of ER on a small number upper-beginner learners at a large New Zealand university. In the first iteration of the tests, the first 2000 words of Mastushita's list were compared to the vocabulary list of the Genki textbook (Banno, Ikeda, Ohno, Shinagawa, & Tokashiki, 2011), which is used for the first two years of Japanese courses at the university. This comparison showed significant differences between the two lists. In light of this, a 200-item test was constructed, sampling every 10th word. Because this was exploratory research, it was not clear what sample rate would be necessary for a Japanese vocabulary levels test, so to avoid the risk of under-sampling and negatively impacting the accuracy of the vocabulary knowledge estimates, a high sample rate was used. In practice, it was found that at the outset of the study, the participants knew on average 88% of the first 1000 most frequent words and 83% of the second 1000 most frequent words. Because even the average score was so high, meaning some participants were scoring close to the maximum score, it was clear the test needed to be extended if it was to be of value to the field. That is, there are likely to be significant numbers of L2 Japanese learners who would not be well-served by the test, as they would already be scoring very close to 100%. In these cases, students will only learn that they should be reading above the 2000-word level. This is not very helpful for learners considering that Matsushita (2012a) estimates that the first 2000 words in Japanese account for only 85% of words in Japanese texts, meaning they would not be able to comprehend most texts they attempted to read.

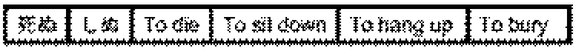
The current version of the test was therefore extended to the 5000-word level, and four forms using different words were constructed. The test used a systematic sample of every 25th word from the first 5000 lexemes of Matsushita's list, for a total of 40 items per 1000-word band, and a total of 200 items for the full test. Using a systematic sample allows these 1000-word bands to be broken down further, if so desired, in order to assess learner vocabulary at the 500- or 250-word level. This may be useful to match learners more closely with graded reader materials, since most published graded readers for L2 Japanese learners set their reading level bands in roughly 300- or 500-word increments.

In some cases, the rigid sampling method selected some items which were readily recognisable abbreviations in English. Items such as these were considered inappropriate for inclusion as scoring correctly does not require the learner to understand the word in Japanese. Four words were skipped: TV, cm, CO, and JR. In these cases, the next item in the frequency list was used instead. For example, if item #2001 on the frequency list was skipped, item #2002 would be used for test form 1. In these cases, the sampled item for all other test forms would also be shuffled down. That is, test form 2 would normally have used item #2002, but instead, item #2003 was used for test form 2, and so on.

Distractor Construction and Item Presentation Format

Current literature detailing distractor construction focusses on English language tests, however, different issues are at play when it comes to testing L2 Japanese learner vocabulary knowledge. The two major points that test creators need to be aware of are the ability of L1 Chinese learners of Japanese to transfer their hanzi knowledge to Japanese kanji and the semantic information present in kanji. In terms of the first point, Matsushita (2012b) reported that Chinese learners' hanzi knowledge is a significant issue when testing L2 Japanese learner vocabulary knowledge. In his poster presentation brief, he reported that there were items that use kanji but that are difficult to understand for L1 Chinese learners of Japanese based on their hanzi knowledge, which impacted test score analysis. In other language contexts too, learners have scored differentially as a result of the impact of transferability of language knowledge from their L1 to their L2 (see, for example, Laufer & McLean, 2016), indicating this is can be a significant issue that test creators should be aware of. While the impact of L1 Chinese learners' hanzi knowledge was taken into consideration when creating these tests, the second point had the greater effect on the design of the distractor options. Japanese kanji are made up of different parts that often, though not always, point to some part of the semantic domain of the kanji and the impact of this is a necessary consideration when designing distractors for multiple choice items. Nation and Webb (2011) give advice about the kinds of distractors that should be used for English vocabulary tests, however for the J-VLT most of the distractors used were semantically plausible options based on components in the kanji of the target word. In addition, distractors were from the same word class as the target Japanese word.

In this test, the items are presented in Japanese in either as kanji with yomigana (pronunciation guides) or as kana (syllabic script), and then four options in English are given on the right, without a non-defining sentence (see Figure 1 below for an example). Learners are asked to circle the correct answer from among the four options given. Although Nation (n.d.) advises using a non-defining sentence for three reasons, it was not considered a necessary inclusion for a Japanese test. First, Japanese word forms always show the part of speech the word belongs to, so cueing that was not needed. Second, the words sampled were distinguishable from homographs and homonyms through either their different kanji or their different yomigana. The third reason of slightly cueing the meaning was not considered necessary due to the semantic cues present in kanji.



死ぬ	しぬ	To die	To sit down	To hang up	To bury

Figure 1: *Sample Test Item*

Reliability and Validity of the Test

Reliability

The reliability of the test was calculated using data collected in early 2021, when the test was administered to third year, first semester students (N = 31) at a large New Zealand university. Reliability was calculated using the Kuder-Richardson 21 formula, and was estimated to be at least .98 (see Table 1 below), indicating very high reliability. In total, the four forms of the test have been administered to more than 100 students in this context during trialling over two years, however, not every administration is covered by consent documentation from student participants to publish their scores. The sample of 31 participants below represents nearly a complete class, are covered by consent documentation, and their results are not dissimilar from other administrations of the test.

Validity

In regard to arguments for the Japanese Vocabulary Levels Test's validity, it should be noted that validation is an ongoing process, in which the more specificity that can be given about the purpose of the test, the learners it can be used with and the context in which the test can be used, the more prospective users can decide if it will be valid for their own contexts (Schmitt, Nation, & Kremmel, 2020). These aspects have been specified earlier in this paper: it has been designed with learning context of tertiary JFL learners in mind, that is, those learning Japanese outside of Japan; and for the purposes of guiding learners to appropriate starting reading levels for ER in

Table 1: *Descriptive Statistics of 200-item Vocabulary Levels Test*

	Maximum score	Minimum score	Mean score	Std. Dev.	KR-21
	(out of 200)	(out of 200)			
N=31	187	78	138.48	30.998	.98

Japanese and to measure any gains in vocabulary that may occur as a result of reading extensively in Japanese. The test's validity for both these purposes is necessarily tentative at this point. Inferences have been made based on the ongoing debate in the L2 English learner vocabulary measurement context, but, as has been explained above, English and Japanese employ very different vocabulary systems. More development in this new field is required, including the development of reading comprehension measures that can support better examination of the relationship between reading comprehension when engaged in ER and vocabulary knowledge for L2 Japanese learners.

In addition, Schmitt, Nation, and Kremmel (2020) argue that at the least, test developers need to also specify how to interpret the test scores and demonstrate that the test scores are reliable. Reliability for this test, shown above, is very high, indicating that if a learner sat the test again a short time after the initial administration, the result should be very similar. Score interpretation relates to appropriate inferences about what the test results demonstrate about learners' L2 vocabulary knowledge. Much of this has been covered above: the test measures learners' recognition of an English equivalent of a Japanese vocabulary item from among four options. Scores cannot be used to infer productive vocabulary knowledge, vocabulary knowledge for listening, or any other type of vocabulary knowledge. That is beyond the scope of the test in its current form. Test results may be interpreted to mean that learners can recognise an English-equivalent meaning of these words. Potentially, it may measure what Laufer and Aviad-Levitzky (2017) term "comprehension vocabulary", however, to validate that would require a frequency-based reading comprehension test, and none have been published so far. Although there is currently a debate about whether multiple-choice tests can assess learners' ability to read at a particular vocabulary coverage level for general reading, this may be less of an issue for ER, where learners of English are advised read at a 98% vocabulary coverage level meaning there are only a couple of words out of every 100 that the reader does not recognise. Furthermore, when engaged in ER learners often read graded reading materials that have plenty of pictures and extra context cues within the text to help readers understand, as well as the additional semantic cues for L2 Japanese readers in the kanji script. This helps to ensure learners have plenty of support to understand the text.

Pedagogical Implications

Several important pedagogical implications arise from the development of these tests. Firstly, this test enables learners, teachers, and researchers to assess learner mastery of the 5000 most frequent Japanese words at the meaning-recognition level for the first time. Importantly, the sampling method means it is possible to adjust the word bands that are assessed to match learner needs in the context of use, such as for matching them to appropriate graded reader levels. This is significantly different to a Vocabulary Size Test, where the result should be taken holistically as an indication of a learners' total vocabulary size in the target language.

Secondly, because it uses the same frequency list as is used in the J-LEX lexical profiling tool, it enables learners to match their vocabulary knowledge to texts at the 95-98% vocabulary level. It means that learners will be able to know if texts they wish to read are likely to be at the right level for them before they start. This is vital to help ensure learners avoid developing foreign language reading anxiety, and have the chance to develop a sense of reading for their own purposes rather than simply using texts for language study. It also means that learners can check the level of any Japanese texts they wish to read for themselves, to know if the text's vocabulary level is too high for them. This means that learners can go beyond the classroom and the currently produced graded reading materials and begin to explore the world of Japanese literature for themselves. In this way, they may be able to develop autonomy and independence in reading in Japanese.

Thirdly, it allows Japanese teachers, researchers, and even learners to measure and track the development of learner vocabulary knowledge of frequent Japanese words. Mastery of frequent words should be an important focus for adult language learners, rather than just the words they encounter in their textbooks, or that are included on the word list of the outdated pre-2010 version of the JLPT.

Finally, it represents a first step towards enabling future research to determine the relationship between vocabulary knowledge level and reading comprehension for Japanese ER and, further, what percentage of words need to be known in order to read extensively in Japanese. In this way, the J-VLT enables the further development of research in the areas of Japanese language teaching, Japanese learner vocabulary acquisition, Japanese extensive reading, and Japanese learner reading comprehension development.

Limitations and Future Directions

As this is the first Japanese Vocabulary Levels Test (J-VLT) to be developed, there are limitations to the design as it currently stands. Additional to the limitations with the format that have been discussed above are the limited number of words assessed. This test only looks at the first 5000 most frequent lexemes, and there are test-takers at the upper beginner level who have scored above 80%, suggesting that learners beyond the lower intermediate level may not be well-served by the information about vocabulary knowledge levels the test provides. There is therefore a need to increase the test's levels to the 10,000 word level so the test may be used with learners beyond the lower intermediate level to both assess learner vocabulary development and to pair learners with suitable texts to read.

There is also a need to develop a meaning-recall version of this test to assess stronger word knowledge, and work is currently underway to complete this process. These types of tests take longer to develop due to the need to account for variation in learner responses given, and the need to assess how much instruction learners need to ensure that they can give enough detail for their answer to be marked correct. Web-based versions of both the translation and multiple-choice J-VLT are also needed to improve convenience.

This paper simply reports on the major considerations when developing this test and its reliability and validity for these learners. Although it argues for the test's validity for the context and purpose of use, it has not yet undergone rigorous statistical analysis. It is intended that such analysis will be carried out in the near future.

This test is currently reliable and valid for one context of use: a large New Zealand university. Work is currently underway to establish its reliability and validity in other universities and countries with different teaching and learning contexts. In addition, the first stage of the development of the translation version has been completed and work is currently underway to refine it.

Conclusion

This paper presents a brief overview of the major considerations addressed when creating the first Japanese Vocabulary Levels Test. It has discussed the lack of such a test for Japanese thus far, and issues that need to be addressed when creating a new vocabulary test. It has outlined the purpose for which the test can be used, the context in which it can be used, and the interpretations of learner vocabulary knowledge that can be made based on the test scores. There are some limitations regarding the test's current form, and collaborative research projects would be welcome in order to further develop the J-VLT in the future.

References

Abe, K. (2016). *Roles of technology and of reading among five self-directed adult learners of Japanese as a foreign language* (Unpublished PhD thesis). The University of Texas at Austin, Texas, USA.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Banno, E., Ikeda, Y., Ohno, Y., Shinagawa, C., & Tokashiki, K. (2011). *Genki: An* *integrated course in elementary Japanese* (2nd ed.). Japan Times.

Banno, E., & Kuroe, R. (2016). Effects of Extensive Reading on Japanese Language Learning. *Proceedings of the Third World Congress on Extensive Reading*, *3*, 1-9.

Day, R., & Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge University Press.

Day, R., & Bamford, J. (2002). Top ten principles for teaching extensive reading. *Reading in a Foreign Language*, *14*(2), 136-141.

Fukumoto, A. (2004). 日本語教育における多読の試み [An experiment in extensive reading in Japanese education]. *Japanese Language and Culture*, *30*, 41–59.

Gyllstad, H., McLean, S., & Stewart, J. (2020). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, *38*(4), 558-579. https://doi.org/10.1177/0265532220979562

Gyllstad, H., Vilkaite, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *International Journal of Applied Linguistics*, *166*(2), 276-303. https://doi.org/10.1075/itl.166.2.04gyl

Hitosugi, C., & Day, R. (2004). Extensive reading in Japanese. *Reading in a Foreign Language*, *16*(1), 20-39.

Horiba, Y., Matsumoto, J., & Suzuki, H., (2006). 日本語学習者の語彙知識の広さと深さ [Breadth and depth of vocabulary knowledge in Japanese as a second language]. 言語科学研究：神田外語大学大学院紀要, *12*, 1-26.

Kageyama, T., & Kishimoto, H. (Eds.) (2016). *Handbook of Japanese lexicon and word formation*. De Gruyter, Inc.

Kess, J. F., & Miyamoto, T. (2000). *The Japanese mental lexicon: Psycholinguistic studies of kana and kanji processing*. John Benjamins Publishing.

Kremmel, B. & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377-392. https://doi.org/10.1080/15434303.2016.1237516

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. & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly*, *13*(3), 202-217. https://doi.org/10.1080/15434303.2016.1210611

Leung, C. (2002). Extensive reading and language learning: A diary study of a beginning learner of Japanese. *Reading in a Foreign Language*, *14*(1), 66-81.

Matsushita, T. (2012a). *In what order should learners learn Japanese vocabulary: A corpus-based approach* (Unpublished PhD thesis). Victoria University of Wellington, Wellington, New Zealand.

Matsushita, T. (2012b). 「日本語を読むための語彙量テスト」の開発 [Development of a vocabulary test for the purpose of reading Japanese]. Poster presented at the meeting of *2012*年日本語教育国際研究大会, Japan.

Matsushita, T. (2014, August). How is the relationship between vocabulary knowledge and comprehension? A case of Japanese. Paper presented at the meeting of the *AILA World Congress 2014*, Brisbane. Retrieved from http://www17408ui.sakura.ne.jp/tatsum/presentation/Matsushita2014_AILA_Vocab-Reading.pdf

McLean, S., Raine, P., Pinchbeck, G., Huston, L., Kim, Y. A., Nishiyama, S., & Ueno, S. (2021). The internal consistency and accuracy of automatically scored written receptive meaning-recall data: a preliminary study. *Vocabulary Learning and Instruction*, *10*(2), 64-81.

McQuillan, J. (2016). What can readers read after graded readers?. *Reading in a Foreign Language*, *28*(1), 63-78.

Miyaji, H. (1969). On the definition of Japanese words and word classes. *Word*, *25*(1-3), 228-244.

Miyaoka, Y., Tamaoka, K., & Sakai, H. (2011). 日本語語彙テストの開発と信頼性 ─中国語を母語とする日本語学習者のデータによるテスト評価─ [Japanese lexical knowledge test and its reliability: Test evaluation using data from native Chinese speakers learning Japanese]. 広島経済大学研究論集, *34*(1), 1-18.

Nation, I.S.P. (1990). *Teaching and Learning vocabulary*. Heinle.

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

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

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

Rothville, K. (2019). *Extensive reading in Japanese: Investigating the effects on vocabulary, reading comprehension, coursework and attitudes towards reading in Japanese and kanji over 27 weeks for upper beginner-level learners* (Unpublished master's thesis). University of Auckland, Auckland, New Zealand.

Rothville, K. (2022). Japanese extensive reading materials: The relationship between the vocabulary of current materials and learners' vocabulary knowledge. *The PanSIG Journal 2021*, 169-177.

Saito, Y., Horwitz, E. K., & Garza, T. J., (1999). Foreign language reading anxiety. *The Modern Language Journal*, *83*(2), 202-218. https://doi.org/10.1111/0026-7902.00016

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development. *Language Teaching*, *53*, 109-120.

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*, 181-203. https://doi.org/10.1017/S027226312000025X

Suganaga Y., & Matsushita T. (2013). *J-LEX Profiler*. Retrieved from http://www17408ui.sakura.ne.jp/

Tabata-Sandom, M., & Macalister, J. (2009). That 'eureka feeling': A case study of extensive reading in Japanese. *New Zealand Studies in Applied Linguistics*, *15*(2), 41–60.

The Japan Foundation Association of International Education. (1994). 日本語能力試験出題基準 [Japanese language proficiency test: test content specifications]. Bonjinsha.

Tamaoka, K. (2014). The Japanese writing system and lexical understanding. *Japanese Language and Literature*, 48(2), 431-471.

Tono, Y., Makawa, K., & Yamazaki, M. (2013). *A frequency dictionary of Japanese: Core vocabulary for learners*. Taylor & Francis.

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*, 454-461.

Webb, S., & Macalister, J. (2013). Is text written for children useful for L2 extensive reading? *TESOL Quarterly*, *47*(2), 300-322. https://doi.org/10.1002/tesq.70

Yamashita, J. (2013). Effects of extensive reading on reading attitudes in a foreign language. *Reading in a Foreign Language*, *25*(2), 248-263.