TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

An Investigation into the Roles of Guessing and Partial Knowledge in the Vocabulary Size Test

November 2022 – Volume 26, Number 3

https://doi.org/10.55593/ej.26103a15

Steven Asquith
Rikkyo University
<stevenasquith@rikkyo.ac.jp>

Abstract

Although an accurate measure of vocabulary size is integral to understanding the proficiency of language learners, the validity of multiple-choice (M/C) vocabulary tests to determine this has been questioned due to users guessing correct answers which inflates scores. In this paper the nature of guessing and partial knowledge used when taking the Vocabulary Size Test (VST) is examined. The analysis evaluates the thought processes of test takers through think-aloud protocols and self-reports. This provides a taxonomy of seven types of guesses used by learners while taking the VST and measures the frequency of their occurrence. Based on qualitative coding, guesses are investigated to determine if they exhibit partial word knowledge which is relevant to the test construct. The findings suggest that both guesses resulting from the use of partial knowledge and random guesses are included in estimates and this has a detrimental effect on test accuracy. The paper concludes that the VST does not provide an accurate measure of the vocabulary necessary for reading due to guessing and the meaning-recognition format. It also suggests that the role of partial knowledge should be considered when producing more sophisticated vocabulary tests in the future.

This study investigates the nature of guessing and partial knowledge used in multiple-choice vocabulary tests through an analysis of the Vocabulary Size Test (VST). The research stems from a discussion relating to the specifications of the VST, which encourage guessing and the use of partial knowledge in order for test takers to get "as much credit as possible for what they know, even if this knowledge is incomplete" (Nation, 2012). Schmitt, Schmitt, & Clapham (2001) however suggest that encouraging guessing in multiple-choice vocabulary tests creates "a certain level of ambiguity" (p. 79) as to how this instruction is interpreted by examinees. Others contend that including guessing in scoring results in vocabulary size being considerably overestimated because test takers receive credit for random correct guesses (Stewart, 2014; Mclean, Kramer & Stewart, 2015). Recent research findings have provided strong evidence that the VST is unable to accurately determine a lexical mastery level for reading (Schmitt, Nation & Kremmel, 2020; Stoeckel, Mclean & Nation, 2020). In addition to guessing, both theory and research strongly suggest that this is because meaning-recognition tests poorly capture the type of lexical knowledge that can be employed when reading relative to a meaning-

recall test (Stoeckel, et.al. 2019; Zhang & Zhang 2020). This study further investigates these issues by using think-aloud protocols and self-reports to categorise and quantify the different types of guesses occurring in the multiple-choice VST.

Guessing in the Vocabulary Size Test

The multiple-choice VST has been a popular resource which is used by learners, educators, and researchers alike as a convenient measure of vocabulary size (Beglar, 2010; Nation & Coxhead, 2014; Elgort, 2013). The VST is available in multiple formats, including paper-based 14,000-word and 20,000-word monolingual versions, numerous bilingual versions, as well as a web-based tool (https://my.vocabularysize.com/). The VST uses a four-choice, multiple-choice format, as shown in Example 1 below, to estimate English learners' receptive vocabulary size.

11. BUTLER: They have a **butler**.

a) man servant
b) machine for cutting up trees
c) private teacher
d) cool dark room under the house

Example 1. An item from the VST

The 14,000-word version adapted for use in this study samples 10 items from each thousand-word band of Nation's BNC list (2012) and contains 140 items given in word frequency order. Results are calculated by multiplying the total score by 100, so that a score of 31 correct items indicates a vocabulary size of 3,100 words. Each item is provided a non-defining context typical of its most frequent environment, so that for example the item *instance* is in the context of the phrase *for instance,* its most common usage. Distractors have been designed so that they do not share core elements of meaning with the correct answer. Therefore, the item testing knowledge of *azalea* simply requires a learner to recognise that it is a plant, rather than knowing specific details (Coxhead, Nation & Sim, 2015).

The purpose of the VST is described in its specifications as to "assess the written word form, the form-meaning connection, and to a lesser degree concept knowledge" (Nation, 2012). The test aims to provide an estimate of the vocabulary knowledge necessary for reading and the test taker is required to recognize a word's meaning by choosing it from a group of distractors, rather than recalling it from memory alone. Coxhead, Nation, & Sim (2015) stated the opinion that the easier requirement of word recognition rather than recall more closely reflects the support learners get during reading from context and background knowledge, and thus, the VST's specifications state that it offers a "slightly generous" estimate of the vocabulary knowledge necessary for reading (Nation, 2012). This view is echoed by Laufer & Aviad-Levitzky (2017) who concluded that both meaning recall tests and meaning recognition tests are "good predictors of reading ability" (p.739). Gyllstad, Vilkaite & Schmitt (2015,) however, assert that the more challenging stipulation of meaning recall is necessary to measure the vocabulary knowledge needed for "fluent reading" (p. 284), stating that the VST is a vocabulary knowledge test not an "inferencing" test**.** Others concur, pointing out that multiple-choice options are not available in the natural act of reading (Stewart 2014). Recently, the evidence against multiple-choice meaning recognition formats being able to accurately indicate reading proficiency has been mounting and many rigorous data-based studies (McLean, Stewart & Batty 2020; Jeon & Yamashita 2014) corroborate this assertion. In fact, the VST designer Nation, now also appears to agree that multiple-choice formatted tests overestimate lexical knowledge (Stoeckel, McLean & Nation, 2020).

Random guessing

The inclusion of guesses in scoring is one of several concerns raised about the accuracy of the VST (Stoeckel, McLean & Nation 2020; Stewart, 2014; Mclean, Kramer & Stewart, 2015). Including guesses in scores results in vocabulary size being overestimated because test takers receive credit for unknown items. This problem is also compounded because each item is representative of a 1,000-word band (Stewart, 2014) and therefore each guess has an adverse influence on the overall vocabulary size estimate (Gyllstad, Vilkaite & Schmitt, 2015). This relates to the VST designers' recommendations that all test takers, including lower levels, should take the whole test, there should be no correction for guessing, and there should be no 'don't know' (DK) option to opt out of answering unknown items (Nation, 2012; Coxhead, Nation & Sim, 2014). These guidelines were based on research findings that even lower-proficiency learners know many low frequency words (Nyugen & Nation, 2011). Furthermore, test takers are encouraged to make informed guesses as these could reflect "sub-conscious" word knowledge, or words that are partially learned (Nation, 2012). Evidence suggests that repetition of encounters with words affects incidental vocabulary learning and that this seems to be incremental as various aspects of words are learned (Webb, 2007). By encouraging learners to make informed guesses the test guidelines aimed to account for such partially learned vocabulary knowledge that may be useful when reading words in a natural context.

The VST's guidelines, however, have led many to observe that if a test taker answers an item on a four-choice, multiple-choice test then they have a 25% probability of getting the correct answer purely through chance. In the case of the 14,000-word, 140-item VST, this would mean that a vocabulary estimate of 3,500 words should be "a near-minimum baseline estimate" (Stewart, 2014, p. 272) for all learners based upon the chance of randomly choosing a correct answer. Learners with small vocabularies, who might have no knowledge of words in large sections of the test, could therefore be expected to amount considerable overestimates of their total vocabulary size if these random guesses are included in scores.

Research conducted into the multiple-choice formatted VLT supports this. Stewart and White (2011) found that guessing inflates VLT test scores by 16 to 17 points when 60% of words are known and Kamimoto (2008) reported a 45% overestimate of scores at the 3000-word level. Mclean, Kramer and Stewart (2015, p. 33) investigated the minimum scores lower-level learners might expect to receive on the VST through random guesses using a three-parameter logistic model. They found that the lowest level learners would be expected to get a score of 2.32 out of ten through random guessing unrelated to vocabulary knowledge, and that this would constitute the "bulk of the score" for an average level Japanese university student. More recent studies have shown that when compared to more difficult and more accurate meaning recall based tests - in which students recall the meaning of a target word from memory without the aid of prompts and accurate guessing is minimised - that multiple choice tests provided scores which were higher by 28.3% with English options (Stoeckel et al., 2019), and 41.9% (Stoeckel et al., 2019), 47.5% Stoeckel & Sukigara, 2018, and 62.7% (Gyllstad et al., 2019) with options in the L1. Such findings suggest guessing and test format have a sizable negative effect upon multiple choice test accuracy.

The extent of guessing is likely to vary among test takers and be influenced by affective factors such as less motivated test takers guessing or skipping questions when an answer is not immediately apparent. Individuals, taking the VST under personal supervision, scored double compared to a group-administered test (Nation, 2012). In excessively difficult tests, as in the case of lower-level learners taking the whole 140 item VST, students may lose motivation and answer items without proper attention. Nation & Coxhead (2014) suggest that the VST is more accurate when word frequencies are mixed throughout the test, rather than progressing from

higher to lower. Time limits could also increase random guessing as less proficient learners may rush to finish the test quickly.

Limiting guessing

The DK option although not included in the paper-based versions of the VST is included in the online test. However, Nation (2012) and Zhang (2013) express reservations about the options use. Nation (2012) states that a DK option is not included in the VST because it discourages "informed guessing". Zhang (2013, p. 808) concluded that having a DK option "slightly improved reliability" but reduced both random guesses and guesses as a result of partial knowledge, and thus lowered the average overall scores. As the decision to use a DK option is subjective it can add an extra layer of ambiguity to scoring (Bennett & Stoeckel 2012), so that students who use the DK option a lot get a lower score than those who use it a little regardless of level. Stoeckel, Bennett, & Mclean (2016) explored the relationship between vocabulary knowledge, test scores, and estimates of reliability for the VST with and without the DK option using simulated data. Based on the findings, they hypothesized that the DK option should not be used as it introduces sources of unnecessary variance unrelated to vocabulary knowledge. Although the DK option is not the primary focus of this study, it is useful to measure the point at which a test taker differentiates a known item from a guess. As this point is based upon a subjective judgement by an individual, with Zhang (2013) and Stoeckel, Bennett, & Mclean (2016) noting a large degree of variation in the DK option's use, it would be unwise to judge this as a definitive measure of the threshold at which a word is known. However, as a vantage point from which to view guessing behavior, it is a useful place to begin.

Informed Guessing

Guessing in a M/C test involves a process of matching the information given in a test item, with the lexical representations integrated in the memory structures of the test taker. Then, based upon the web of connections made between the test item and these lexical representations, whether correct or incorrect, an answer is chosen. This logically means that higher-level learners, with a larger store of lexical representations are likely to be able to make more numerous connections and theoretically, we might assume that they would be able to guess more skillfully. Not only would they be able to connect known words, but also word parts, similar sounding words, or potentially, even to draw from words stored subconsciously. Connecting word parts or similar sounding words to guess unfamiliar or partially known vocabulary are important skills which need to be explicitly taught and encouraged during reading practice (Nation, 2013). Although, the multiple-choice format does not mimic the natural act of reading, it does prompt learners to try to guess unfamiliar or partially known vocabulary by connecting items to their lexical resources. Such behavior could have potential benefits for learners' reading.

Although research on guessing in multiple choice vocabulary tests has mostly focused on the negative effects of random guesses upon accuracy, a few studies have looked more closely at the nature of guessing. Probably the most detailed of these, was conducted by Kamimoto (2008) on the bilingual English/Japanese version of the VLT. As in the current study, think-aloud protocol were used to investigate the types of guesses occurring. Through an analysis of five test takers, the frequency of 12 different guessing categories, of which; elimination, blind (random) guessing, partial knowledge, loanwords and spelling were the most common, were identified and counted. Kamimoto (2008) concluded that even lower-level students can gain substantial score increases through guessing and that they use a range of guessing strategies. More recently Gyllstad, Vilkaite & Schmitt (2015) used post-test interviews to establish what learners knew about guessed items in the VST. In addition to blind (random) guessing, they recorded the strategies of inferring the meaning through either a word family member, a similar

word in the test item, or the context of the sentence; and elimination and association, whereby learners deduce the answer through their knowledge, both correct and incorrect, of the test item and distractors. They noted that the most commonly used guessing strategy by far was elimination and association. Random guessing was judged to be used infrequently, only when, and if, all other strategies had failed, and as it was not often successful in their data sample, did not appear to distort scores. MacDonald and Asaba (2015) also found that test takers only used random guessing as a last resort and that guesses were based upon a range of both correct and incorrect partial knowledge. Although, Gyllstad, Vilkaite & Schmitt (2015, p. 301) judged that inferring the meaning of an item through word family members probably does draw on partial knowledge that is useful for communication, they concluded, that the other guessing strategies "must be seen as undesirable and construct irrelevant". This view has taken even greater precedence as the evidence supporting meaning-recall formats rather the meaning recognition (M/C) has mounted (Stoeckel, et.al. 2019; Zhang & Zhang 2020). This research aims to explore more deeply the types of guessing strategies used by learners taking the VST to provide a better understanding of their nature and quantity.

Research question

The multiple-choice format is ubiquitous throughout vocabulary assessment, primarily because of its ease of use, but many researchers consider it problematic due to its meaning-recognition format and the possibility of guessing the correct answer (Stewart, 2014; Mclean, Kramer & Stewart, 2015; Gyllstad, Vilkaite & Schmitt, 2015). Others, however, suggest that guesses may demonstrate partial knowledge, which otherwise would not be factored into scoring. To investigate the role of guessing in how students answer multiple-choice vocabulary tests, and how these are subsequently scored, this study collected data based upon the following research question.

What are the types and quantities of guesses that learners make when taking the VST?

Methodology

This mixed-methods research used two studies to gather quantitative data on the amount of guessing being used by learners taking the VST, and qualitative data to describe and categorise the types of guessing occurring. By combining the methods, the researcher could not only uncover types of guessing but also quantify the extent to which these guessing types were being used by learners. Quantitative data was collected primarily through the self-reports, while qualitative and quantitative data was uncovered through the think-aloud protocols. This approach is based on the core assumption of mixed-methods research that a combination of quantitative and qualitative methods provides a "more complete understanding of a research problem than either approach alone" (Creswell 2014, p. 33)

Instruments

In both tests, participants used a modified version of the 140-item, 14,000-word VST downloaded from Paul Nation's webpage (Victoria University of Wellington 2022) The test, provided in **Appendix 1**, was modified in the following ways. First, as a trial of the think-aloud test using the whole 140-word VST showed that it would take over two hours, it was decided that it would be better to use half the test by sampling the first five items of each 10-item level. This meant that rather than multiplying the score by 100 to get the vocabulary size estimate, it was necessary to multiply it by 200. Although Nation (2012) and Beglar (2010) state that a scaled down version of the test should work well, recently Gyllstad, Vilkaite & Schmitt, (2015) have expressed concern that each 1,000-word level needs to have a larger sample than 10 to be accurately represented. As the main focus of this study was recording guessing behaviour, and participants' consistent concentration was required, a 70-item test was deemed to be the best

instrument. Additionally, for this reason, the item order was rearranged so that the difficulty was mixed throughout the test (Nation & Coxhead 2014). This may have reduced the number of random guesses due to demotivation in the difficult low frequency levels, especially for the lower-level participants. Finally, the DK option was added to the bottom of each item. This was necessary so that participants could indicate if they believed they knew the item, or if it was a guess. As designated in the VST guidelines all items were attempted. If an item was unknown the participant first circled DK and then attempted to guess the correct distractor. This meant that all items specified DK could be categorized as guesses. Although this design could record what each participant considered to be known or a guess, in the case of overconfident learners, some guesses may still have been indicated as known. Therefore, the number of correct answers for known items is also presented in results.

Think-aloud participants

The ten participants in the think-aloud protocols were from a wide variety of backgrounds, professions, and ages with estimated English proficiencies ranging from upper-intermediate to elementary level. As the think-aloud tests required a time commitment of between 1 to 2 hours the researcher canvased for volunteers from friends, colleagues, students and acquaintances. Participants consented to taking part after being made fully aware of the research procedures, aims and intentions. Four participants provided TOEIC scores (810, 730, 730, 700), seven participants used English regularly in a professional capacity, and two lower-level participants had no formal schooling of English beyond compulsory education. A more detailed description of the participants is provided in **Appendix 2.** Also, importantly all participants were Japanese, meaning that any L1 influences including the use of cognates were uniform.

Think-aloud procedure

Think-aloud protocols, derived from the field of psychology, are useful in tracking participants thought processes as they engage in a task (Heigham & Croker, 2008; Kamimoto, 2008). In this study participants sat the 70-item VST while verbalising their on-going thoughts to a researcher. Participants first read each item, then either circled DK if the word was unknown or chose an option. If the item was deemed unknown, then the participant verbalised their thought processes as they tried to guess the correct option. As it is important to be as unobtrusive as possible, to minimise the time between the thought processes and verbalisations, and also, not to ask leading questions when conducting think-aloud tests (Heigham & Croker, 2008), the researcher limited their interaction with the participants during the test. In order to focus the test taker on whether an item was known, or a guess, the question "Do/Did you know that word?" was asked at some point during each item. If they replied "No", they were then reminded to circle the DK option, if they had not already done so. Other than asking if an item was known, the researcher interjected as little as possible so as not to influence the test process. However, if a participant grew silent for an extended period or if more information was required, probe questions such as: "Why did you choose this one?", "Have you seen this word before?", and "What are you thinking about now?" were used. Also, to help participants to verbalise their thoughts, they were encouraged to use their L1 if necessary. Tests were recorded so any translation issues could be resolved subsequently. Using this procedure all participants were able to give an ongoing commentary throughout the 70-item test. The two lower-level participants however found the test long, difficult and somewhat frustrating. This seemed to validate the decision to create a shortened 70-item VST, rather than using the whole test. Each test took between around one hour for the quickest participant, to just over two hours for the slowest.

Using the think-aloud procedure data was collected and coded from a total of ten participants using a descriptive coding approach to index and categorise the data (Saldana 2011, p. 113).

Prior to coding four provisional categories were anticipated based on the results of Gyllstad, Vilkaite & Schmitt, (2015) and Kamimoto (2008). These were: distractor elimination - or choosing the most likely option through a process of elimination; word parts - or using knowledge of prefixes, suffixes, or word parts to guess the meaning; polysemy - guessing the meaning of an item based upon a different meaning of the same word; and random guessing, choosing the option based purely upon chance, with no knowledge or strategy used. Also, it soon became clear that participants used similarities between known words and the distractors, such as the known word mystery and the distractor *mystique*, to guess words, and this was coded as similar words. The sixth and seventh categories however emerged directly from the data. These were distractor-triggered responses, where the participant initially claimed not to know the word but upon reading the distractors realised that they already knew the word, and semantic sense, where a vague sense or 'memory' of a word was used to guess the meaning based upon the distractors. Also, known items, in which a DK option was not indicated were coded accordingly. As there was some overlap between categories; for instance, learners often used knowledge of the singular sense of sol in *soliloquy* (word part) to choose the distractor *speech in the theatre by a character who is alone* (distractor-triggered response), the researcher coded each guess according to which category was felt to be the best fit, in this case, word part. The transcript samples in **Appendix 3** provide greater clarity on each of these categorizations.

Self-report test participants

The 13 participants in the self-report test were all active members of an intermediate to upper intermediate level class, which engaged in reading unmodified English texts. All members of the class were Japanese most of whom were retired and enjoyed studying English as a hobby to maintain their proficiency. Of those who had taken a formal qualification, two recorded Pre-1st Level EIKEN (2022), five recorded 2nd level EIKEN (2022), with TOEIC (2022) scores of 900, 885, 850, 700 and 650, also recorded. However, most of the students had not achieved a formal qualification recently and these scores may be misleading. Overall, the majority of the class students were able to communicate smoothly and fluently in English and this corresponds with the level assessment of intermediate to upper-intermediate made by the researcher. Students were fully informed and consented to participate in the research.

Self-report test procedure

In the second stage of testing, a group of 13 learners took the 70-item VST before indicating on a check sheet how unknown items were guessed. This was conducted to provide a larger, if still somewhat limited, sample size to measure the quantities of guesses. The check sheet was designed based upon categories that emerged during analysis of the think-aloud tests. These categories were phrased in the check sheet as: 100% no idea, just a sense or feeling, a similar English word, a similar Japanese word, part of the word, from the other options, and other. Participants were first asked to complete the VST using the DK option. They were informed in both the test instructions and verbally not to guess, but to indicate DK for words they had not seen before. After all participants had completed the 70 test items, they then went back and guessed the items marked as DK. While answering these, participants recorded how their guesses were made on the check sheet. Individuals, in many cases used more than one strategy to make a guess. For instance, in the case of the item *devious*, they may decide that the prefix *'de'* denotes something negative and accordingly eliminate two positive options as incorrect. In this case both categories 'part of the word' and 'from the other options' would have been indicated on the check sheet. Although initially, a few individuals experienced difficulties in understanding what was required, these were quickly resolved. It was clear however, that categorizing guesses into distinct categories in this way was difficult and very subjective.

Therefore, this test was deemed less precise than the think-aloud tests and used purely to record the quantities of known items, the number of guesses, and their associated scores.

Analysis and results

Quantities and types of guesses made by students while taking the VST, and the factors resulting in correct guesses were evaluated. Guesses were defined as any answer given where DK has been specified by the participant as this enables the researcher to distinguish which answers were perceived as unknown.

Quantities of guessing occurring in the VST

Firstly, the results from the think-aloud protocols and self-reports showing the sizable effect guesses have on VST scores are presented. Tables 1 and 2 below, present the scores of learners on known and unknown items. If a test taker indicated DK on the test sheet, an item was categorised as unknown and therefore a guess. Correct answers on unknown items were categorised as correct guesses (CG). The tables are arranged by total scores achieved on the VST. This shows the quantity and efficacy of guessing, and how this is linked to the VST scores.

Table 1. The quantities of known items, guesses, and corresponding scores for the think-aloud VST

		Items Marked as Known		Items Marked Don't Know	
Learner	VST Score (k=70)	Total % of Test	Correctly Answered (% of Score)	Total (% of Test)	Correctly Answered (% of Score)
A	52	38 (54%)	37 (71%)	32 (46%)	15 (29%)
B	48	28 (40%)	28 (58%)	42 (60%)	20 (42%)
C	47	36 (51%)	35 (74%)	34 (49%)	12 (25%)
D	46	34 (49%)	34 (74%)	36 (51%)	12 (26%)
E	45	34 (49%)	29 (64%)	36 (51%)	16 (36%
F	44	24 (34%)	21 (47%)	46 (66%)	23 (52%)
G	42	41 (58%)	35 (87%)	29 (41%)	7 (17%)
H	39	36 (51%)	29 (74%)	34 (49%	10 (27%)
I	30	12 (17%)	11 (37%)	58 (83%)	19 (63%)
J	27	20 (29%)	14 (52%)	50 (71%)	13 (48%)

Note: The total number of test items was 70

Table 2. the quantities of known items, guesses, and corresponding scores for the self-report VST

Learner	VST Score (k=70)	Items Marked as Known		Items Marked Don't Know	
		Total % of Test	Correctly Answered (% of Score)	Total (% of Test)	Correctly Answered (% of Score)
K	60	45 (64%)	43 (72%)	25 (36%)	17 (28%)
L	57	61 (87%)	52 (88%)	9 (13%)	5 (9%)
M	54	53 (71%)	43 (80%)	17 (24%)	11 (20%)
N	53	57 (81%)	52 (98%)	13 (19%)	1 (2%)
O	50	52 (74%)	42 (84%	18 (26%)	8 (16%)
P	50	48 (67%)	42 (84%)	22 (31%)	8 (16%)
G	49	58 (83%)	46 (93%)	12 (17%)	3 (6%)
R	48	39 (56%)	36 (75%)	31 (44%)	12 (25%)
S	45	30 (43%)	27 (60%)	40 (57%	18 (40%)
T	45	41(59%)	34 (76%)	29 (41%)	11 (24%)
U	42	40 (57%)	31 (77%)	30 (43%)	11 (26%)
V	41	25 (36%)	21 (51%)	45 (64%)	20 (49%)

Note: The total number of test items was 70

The considerable extent to which the inclusion of guessing increases scores is evident in Table 1 and Table 2. This can be seen clearly in the difference between the scores on known items and the total scores which include correct guesses. These correct guesses increased the total scores in the think-aloud tests from between a minimum of seven points to a maximum of 23. In the self-report test, this ranges from one to 20. If the average increase of score through correct guesses over all 23 tests is calculated, it comes to 12.5 items. Considering that each item is equivalent to 200 words in this 70-item test, then 12.5 items would equate to an increase of 2500 words through guesses. This is a sizable effect and amounts to 27.1% of correct answers over all 23 participants being categorised as guesses. In the case of the test taker who correctly guessed 23 items, the vocabulary size estimate would increase by 4600 words, from 4200 words to 8800, or over double the score if guesses were omitted. Furthermore, the largest increases in vocabulary size estimates in relation to total scores, occurred in the lowest level participants. This demonstrates the sizable influence guessing can have on scores and shows how this is amplified through multiplying the raw scores by 200 to create an estimate. Although, this effect would exert less influence in the full 140-item VST version, which has 10 words from each frequency level and multiplies scores by 100, it reinforces Gyllstad, Vilkaite & Schmitt's (2015) assertion that this method of calculating estimates puts a considerable emphasis on the accuracy of each individual item.

In addition to the influence of guessing upon scores, the table above also gives an indication of how the threshold at which test takers deem a word known or unknown varies between individuals. As expected, the data shows that the two lower-level learners with much smaller vocabularies chose DK more often. However, the number of DKs fluctuated greatly between individuals. For instance, looking at the 13 test subjects who scored between 41 and 50, the number of total DKs ranged from 12 to 46. During observation it was evident that some

participants were much less confident of what they knew, or naturally conservative in their appraisals. Others were overconfident and likely to guess whether a DK option was present or not. This is evident, particularly in the self-report test data, where many "known" items were incorrect, supporting the observation that unknown items may not represent all the guesses occurring in the VST. Similar variability in the use of a DK option was found by Zhang (2013), Stoeckel et al 2019, and Stoeckel, Bennett, & Mclean (2016). The variation between participants in determining what is known or unknown suggests that controls for guessing such as a DK option may simply add an extra layer of subjectivity to multiple-choice tests. This also has implications for the validity of self-appraisal type tests such as the Yes/No Test (Meara & Buxton 1987) which use this decision as the primary method of measurement. The effectiveness of guessing as a strategy for getting the correct answer can be seen in the proportions of correct to incorrect guesses in Figure 1.

Figure 1 shows that the overall effectiveness of guessing as a strategy in the think-aloud and self-report tests was far higher than had it been purely as a result of a one in four chance. Although one might expect random guesses in a four-option multiple-choice test to be correct approximately 25% of the time (Stewart, 2014), the proportion of correct to incorrect guesses was 39.2%. This, however, is a considerable oversimplification. As anticipated, the data suggests that learners with larger vocabularies and greater lexical resources are more adept at using partial knowledge to make informed guesses. For instance, test takers that scored over 50 points, equating to a 10,000-word vocabulary estimate, could guess the correct answer 47.8% of the time. In contrast however, learners with smaller vocabularies were less able to make correct guesses. Test takers that had a total score of less than 40 were able to guess the correct answer only 30.5% of the time, far closer to the 25% suggested by Stewart (2014), and consistent with the findings of Mclean, Kramer and Stewart (2015) which showed the bulk of lower-level learners' scores to be made up of random guesses.

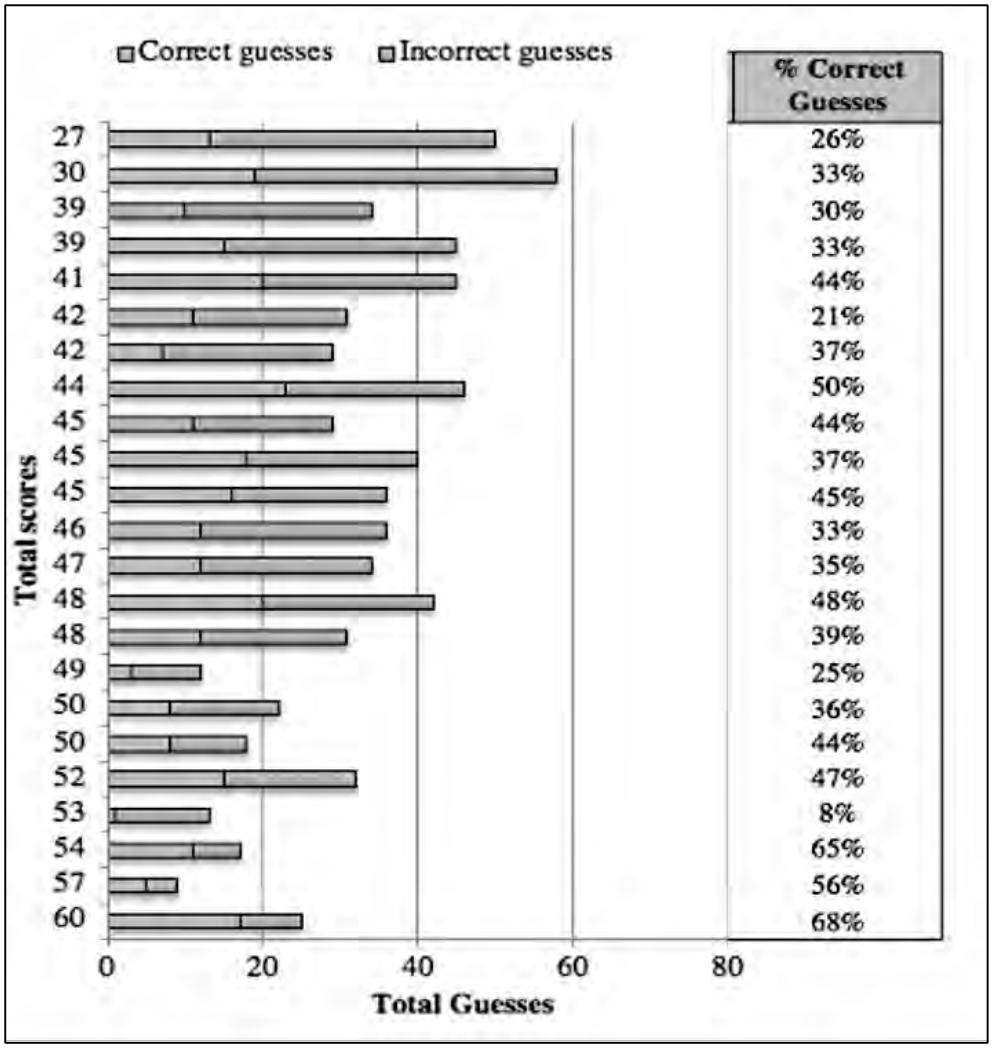

Figure 1. The Proportions of Correct to Incorrect Guesses over all 23 Participants, and the Percentages of Correct Guesses According to Total Score

Correct random guesses.

Only the think-aloud tests were used to identify types of guesses as these provided much more detailed data. When participants indicated no knowledge of an item and none was displayed, the answer was classified as a random guess (RG). If some knowledge was demonstrated, or any strategy employed, this was categorized as an informed guess. The effect on scores of random guesses and informed guesses in the think-aloud tests is presented in Figure 2.

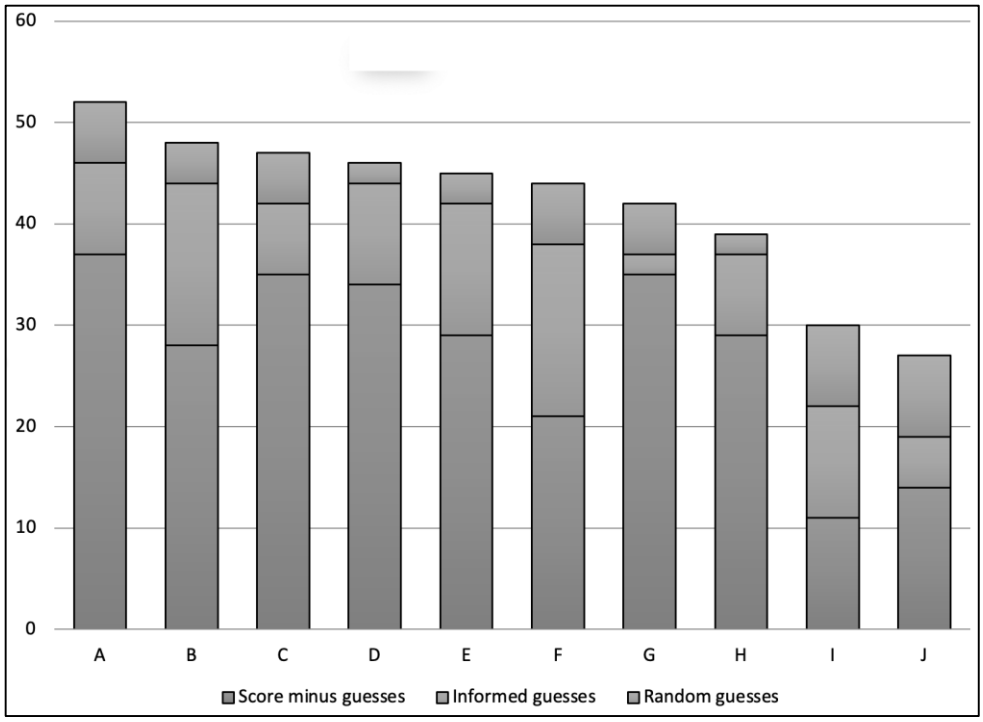

Figure 2. The Quantities of Informed Guesses and Random Guesses in the 10 Think-aloud Tests

Figure 2 shows that all 10 participants labeled A to J who took the VST using the think-aloud procedure had overestimated scores due to random guesses. This ranged from one item, adding 200 words to the vocabulary size estimate, to eight items, which added 1600 words. In the two lowest scorers this represented a sizable overestimate, adding approximately a third to the total score. Even the highest scoring person, who was likely to be good at making informed guesses, had five correct answers resulting from pure chance. This inflated their vocabulary size estimate from 9,400 to 10,400. As this figure does not include guesses in which some strategy was used, concern as to the effect of random guessing, particularly in lower-level test takers who guess more frequently, is clearly justifiable. However, overall, informed guesses were more frequently used than random guesses. In total of 147 correct guesses, 43 were categorised as random and 104 as informed. These informed guesses represent a range of different knowledge and strategies, and as such, to understand their relevance to vocabulary size estimates it is necessary to examine them in greater detail.

Correct informed guesses.

At the heart of this debate is the view that informed guesses may draw upon partial or "sub-conscious knowledge" (Nation, 2012), and that this knowledge is relevant to the purpose of the VST in providing an estimate of the vocabulary knowledge necessary for reading. In the think-aloud tests, in total 104 correct answers were classified as informed guesses. Table 3 below provides a brief summary of each guessing type including an example, the number of occurrences, and if partial knowledge was evident. A more detailed explanation of each type is provided in Appendix 3**.** Guessing categories seemed to emerge from the participants' personal interpretations, and as such can be considered a starting point for further investigation. They also add greater detail to the testing strategies identified by Gyllstad, Vilkaite & Schmitt, (2015). It is hoped that these categorisations stimulate further investigation into this interesting and under-researched area of vocabulary test design. Further details about the quantities of guessing types are shown in **Table 4**.

Table 3. A taxonomy of the guesses found through the think-aloud protocols in the VST

Guessing Category	Partial knowledge	Description	Example *Test items are **highlighted** and correct distractors are underlined. Commentary in *italic*	Quantity *Total guesses ** % of total cor-rect answers (420)
Distractor-Triggered Responses	In most cases yes	Although initially specified as unknown, when encountered in a distractor the test taker realised that they had some knowledge of the item	*Trying to read the item* "Hmm… Ro-bel. Hmm… **Ru-ble**. *(Scanning through the distractors and finding 'Russian money')* Ah, I got it, this one! Like Japanese ru-be-ru".	25 6%
Similar words	Yes	Using similarities between known words and items to guess	*Reading* "Communiku?, **Communiqué,** perhaps? OK, this one… because announcement is similar to communicate".	16 3.8%
Word part	Yes	Using a suffix, prefix, or word part to guess the correct answer	"so so **soliloquy**, soli hmm… soli means…single…" *reading through the distractors until encountering 'alone'*. "Ah, this one".	17 4%
Polysemy	Yes	Guessing the correct answer through a different meaning of the word	*Reading* "they gave us some **accessories**…" *long pause* "hmm, extra pieces? But.. don't we use accessories for like.. earrings..?" *Reading again* "Ok, in that case, this one".	11 2.6%
Semantic Sense	Negligible	Correctly guessing based upon a vague sense of the word. Usually used together with distractor elimination	"Hmm, **Nun**, I don't know. I'll try to use my imagination" *After a long pause* "OK, this one… It feels religious … Maybe I have heard it before".	29 6.9%
Distractor elimination	None	Purely exploiting the test format by using a process of elimination without any partial knowledge demonstrated	**"Threshold,** I don't know!" *Going through the distractors* "Flag?.. Roof?.. Point where something changes? Cost of borrowing money? Ok, it's not this one, it's commission. OK…, these are concrete and this is abstract. I'll try the abstract one!"	12 2.9%
Random Guess	None	No partial knowledge or guessing strategy shown	"This one perhaps, or this one… Sorry, I have no idea".	43 10.2%

Table 4. The quantity of guesses in each think-aloud test expressed as a percentage of the total score

Total score	Known items	Correct guesses	Random guesses	Distractor triggered responses	Similar words	Word parts	Polysemy	Semantic sense	Distractor Elimination
52	71.2%	28.8%	9.6%	5.8%	3.8%	1.9%	1.9%	7.7%	0%
48	58.3%	41.7%	6.3%	2.1%	12.5%	6.3%	2.1%	16.7%	0%
47	74.4%	25.5%	8.5%	2.1%	2.1%	2.1%	2.1%	4.2%	6.3%
46	73.9%	26.1%	2.2%	4.4%	2.2%	6.5%	4.4%	6.5%	4.4%
45	64.4%	35.6%	4.5%	6.7%	0%	8.9%	2.2%	8.9%	4.5%
44	47.7%	52.3%	11.4%	11.4%	9.1%	6.8%	2.3%	6.8%	4.5%
42	83.3%	16.7%	11.9%	0%	2.4%	0%	2.4%	0%	0%
39	74.4%	25.6%	5.1%	10.3%	2.6%	2.6%	0%	5.1%	2.6%
30	36.7%	63.3%	26.7%	16.7%	0%	0%	6.6%	10%	3.3%
27	51.9%	48.1%	29.6%	3.7%	0%	3.7%	3.7%	0%	3.7%
420	**65%**	**35%**	**10.2%**	**6%**	**3.8%**	**4%**	**2.6%**	**6.9%**	**2.9%**

Distractor-triggered responses

In many cases these guesses showed considerable partial knowledge of the item, but it is contentious whether this would be useful while reading naturally. Successful guesses resulting from distractor-triggered responses occurred on 25 occasions and made up 6% of the total scores. These are items where the participant initially claimed not to know the word, but upon reading the distractors realised that they knew the correct answer. A good example is the item *rouble*. This item was guessed five out of 10 times. In this case, the individual had difficulty reading the word and assumed that it was unknown. However, after reading the distractor, *Russian money*, participants realised that the Japanese word *ruberu* was in fact the same word. A similar process occurred with the word puritan, which is also used in Japanese. Daulton (2007) expounds the usefulness of these English loan words in Japanese describing them as a "built in lexicon", while Elgort (2013), in her study of a bilingual version of the VST, found that such cognates are significantly more likely to be answered correctly by test takers. Elgort (2013) emphasized the importance of including these cognates to an equal proportion that is in the L1 in order to produce accurate vocabulary size estimates. Laufer and McLean (2016) also noted this "loanword bias" (p. 215) and proposed that future research should look at how loanwords are distributed in word frequency models. Other distractor-triggered responses included, the items *alum*, *atop,* and *egalitarian* which were triggered by the answers, through either the obvious connections between *alum* and *aluminium* or *atop* and *at the top of*, or the less explicit *equal* and *egalitarian*. On other occasions, participants suggested that seeing the distractor triggered recall of a word they knew but had forgotten. While coding these guesses, the researcher typically looked for an "Aha" moment in which the participant would stop on a distractor and immediately signal that it was the correct answer. Although distractors do not occur in natural reading, they may provide a "conceptual starting point" (Beglar & Nemoto 2014, p. 4) to recall a meaning, and it might be argued that this is what readers gain through context. Laufer & Aviad-Levitzky (2017) describe such vocabulary as "comprehension vocabulary" and contend that such vocabulary, which can be guessed from cues or context, mean that vocabulary recognition tests may be closer indicators of reading ability than recall tests. This view, however, is not well supported by the literature (Stoeckel, et.al. 2019; Zhang

& Zhang 2020). An example of such guessing is that a Japanese learner would be more likely to recognise the meaning of *rouble* in the context of a financial report, a currency exchange rates table, or a Russian travel guide. Similarly, the item *atop* might be guessed easily within context. Although in many cases learners had partial knowledge of these distracter-triggered responses, they were unable to recall the meaning from a decontextualized test item alone. However, once provided a cue through the distractor they were usually successful in choosing the correct answer.

Similar words

The strategy of using similarities with other known words was effective in producing a correct answer on 16 occasions or 3.8% of the total scores. This strategy was also noted by Gyllstad, Vilkaite & Schmitt (2015). A good example of this is the item *mystique,* which was guessed correctly six out of 10 times. Participants, based upon the similarity between mystery and *mystique,* could connect this to '*the secret way*' in the distractor. This illustrates how web-like connections interact between test takers' lexicons and distractors to form answers. Other examples were the use of similar words such as *communiqué* and *communication,* and *jovial* and *joy.* The correct answer to communiqué was guessed on four out of six occasions showing how effective this strategy can be. However, this strategy was also open to misinterpretation. For instance, the similarity between *weir* and *weird* led seven out of 10 participants to choose the option *a person who behaves strangely*. Overall, however, it seems that this type of guessing through similarities would be a valid part of recognising unfamiliar words.

Word parts

The use of word parts, including suffixes and prefixes, to learn and understand words in context is a valuable and recognised reading skill (Nation, 2013). As expected, this featured in the guessing strategies used, with 17 correct answers (4% of total scores) being deduced from part of the word. The most common instance of this was knowledge of *sol,* as meaning alone, in *soliloquy*, which led participants to correctly guess the answer on five out of eight occasions. Among others, *monologue*, *excrete* and *counterclaim* were also guessed in this way. Counterclaim was notable as several participants guessed the word through knowledge of *counter* but were then confused by *claim.* This was because the Japanese equivalent loan word *kure-mu*, also carries the meaning of making a complaint or asking for money to be returned when receiving bad products or services at a shop. This meant that participants were confused by the second distractor '*a request for a shop to take things back with faults'* and highlights the dangers of assuming that loan word meanings conform exactly to L2 equivalents. Word parts can provide useful clues for learners to identify multiple unfamiliar lexical items, and as such, they are part of the construct of vocabulary knowledge.

Polysemy

The capability of multiple-choice tests to target a specific meaning is a useful strength. By focusing on specific meanings, they can target what is specifically known about a word, an advantage over some other test formats. For instance, a test designer might target a more difficult meaning of a word such as a jelly *sets,* rather than a commonly known one such as a train *set*. In a Yes/No test the test taker would likely indicate that they knew a word based upon its easiest meaning. When participating in the think-aloud protocol, test takers could deduce an unknown meaning from a known one on 11 occasions (2.6% of total scores). For instance, the Japanese loan word for *accessory* only includes the definition of jewelry, rather than *extra pieces* as in the distractor. Also, *premier,* as in *the head of government,* was deduced from the superlative notion of the '*Premier*' League of English football. Unlike distractor-triggered responses in which recall of the meaning was prompted by the distractor, participants were

often surprised when the expected definition was not found amongst the distractors. They then had to return to the knowledge they had of the different meaning to deduce an answer. This means that the effectiveness of this strategy to understand an item is likely to be largely determined by how closely the different meanings of the word are linked semantically. This factor would also potentially govern if the word could be understood while reading.

Semantic sense

Unlike the other forms of educated guesses that have clear links between word knowledge and correct choice, the 29 answers (6.9% of total scores) classified as semantic sense are more vaguely connected to general connotations, categories, or concepts related to the words. These are correct answers that the test taker indicated some sense of the word, such as, it is positive or negative, it is not a physical object, or simply some feeling. For instance, several test takers suggested that the word *eclipse* had some connection to space or the field of astronomy. Based upon the fact that an eclipse had recently received a lot of news coverage in Japan, there was a good chance that participants had encountered this word recently. We might speculate that this type of guess constitutes the partially learned words or 'sub-conscious knowledge' that Nation (2012) suggests should be included in vocabulary size estimates. However, these guesses were usually made by connecting vague ideas to the distractors, and as such, seem to have little relevance to estimating the vocabulary knowledge necessary for reading.

Distractor elimination

These were items in which a guess was made based upon eliminating distractors that were believed incorrect, and then guessing. These 12 answers were mostly used in combination with some other strategy. A good example is the item *whim*, which six out of 10 participants guessed correctly. The item was presented as below.

13. WHIM: He had lots of **whims.**

    a) old gold coins
    b) female horses
    c) strange ideas with no motive
    d) sore red lumps

Example 2. An item in which distractor elimination was used

As can be seen, three of the distractors are concrete things and only one is an abstract concept. Individuals who guessed the correct answer often said they sensed a *whim* was an abstract concept. However, it is impossible to tell if this was prompted by knowledge of the word, or by simply recognising the semantic odd one out. Similarly, individuals who had a positive semantic sense of an item might eliminate two negative items, and then guess between the two remaining. This type of guessing was different to distractor-triggered responses in which participants appeared to have significant partial knowledge of a word and were usually able to guess correctly. Distractor elimination was usually unsuccessful and had only a negligible effect (2.9%) on the total correct answers. This strategy is based purely on exploiting the test format and therefore should be deemed as entirely negative to test accuracy.

Discussion

Accurately estimating the total vocabulary size of an individual's lexicon from a practical and usable test is an extremely difficult task. Not only are words themselves "slippery customers, with vague boundaries…, fuzzy edges" and semantic information that is extremely difficult to define (Aitchison 2012, p. 54), but vocabulary knowledge exists upon a scale, potentially meaning anything from the minimum knowledge necessary to guess a word's 'meaning' in a natural context, to the full array of forms, meanings, and usages described in such detail by

Nation (2013, p. 49). The decisions made by test designers about the format, the depth of word knowledge measured, the items chosen, and the way estimates are calculated, often have a large influence on the resulting scores and estimates. Interpretation of these estimates therefore needs to be carefully evaluated based upon the features of the specific instrument used until more rigorous and precise tests are developed (Schmitt, Nation & Kremmel, 2020) The VST specifications state that by testing word recognition and including guessing to account for partial word knowledge it offers a "slightly generous" estimate of the vocabulary knowledge necessary for reading (Nation, 2012). However, based upon analysis of the guessing strategies used when taking the VST, it seems that many answers included in size estimates go beyond the scope of even these generous testing parameters. This finding suggests that the VST lacks accuracy in measuring the vocabulary knowledge necessary for reading based upon guessing behavior. Other research has further indicated that there are fundamental flaws in the meaning-recognition (M/C) format which makes it unsuitable to measure fluent reading (Stoeckel, et.al. 2019; Zhang & Zhang 2020).

According to the data collected from the 23 participants who took part in the think-aloud protocol and self-report, guessing considerably inflates VST estimates. These figures presented in Table 2 show that on average 27.1% of the total scores were items that had been specified as unknown and then subsequently guessed as instructed in the VST guidelines. However, several details provide greater insight into this finding. First, the total success of guesses was of a much higher proportion 39.2%, than was likely to be achieved purely by chance. Second, higher scoring learners with theoretically greater lexical resources were more successful guessers than lower scoring learners, and third, participants used a range of different guessing strategies which included random guessing, exploiting the multiple-choice format, and using partial word knowledge. The qualitative findings of the 10 think-aloud tests help us to understand these details further.

By analysing data from the participants collected during the think-aloud procedure, seven types of guesses were identified. These described in Table 3 are random guesses, distractor-triggered responses, similar words, word parts, polysemy, semantic sense and distractor elimination. Of these, random guesses, which accounted for 10.2% of all correct answers in the think-aloud tests were clearly construct irrelevant. As in the findings of Mclean, Kramer and Stewart (2015) these were most problematic in lower scoring students, and they constituted over 25% of the total scores of the two lowest scoring participants. These totals were far lower in higher scoring participants suggesting that there is less measurement error in the observed scores of the more proficient participants. However, a serious amount of measurement error is introduced by requiring lower-level learners to make guesses on a test which is mostly far beyond their proficiency levels. The 2.9% of answers categorised as *distractor elimination* are also a consequence of the test format and as such irrelevant to vocabulary size estimates. The semantic sense (6.9%) category might also be grouped with distractor elimination as in most cases these were used together to produce a correct answer. Although some of these instances may demonstrate very limited partial knowledge, this is unlikely to be usable in a real context. However, the format of the VST allowed the categories; similar words, word parts, and polysemy, constituting approximately 10% of the total scores, to be used to effectively guess the correct answer. These may be seen as construct relevant types of partial knowledge that provide support when reading naturally in context if one ascribes to Laufer & Aviad-Levitzky's (2017) somewhat disputed concept of "comprehension vocabulary". The final category, distractor-triggered responses (6%), is the most debatable. Clearly, when reading a learner does not have access to distractors and therefore these responses must be irrelevant to measuring the vocabulary knowledge necessary for reading. However, while conducting the think-aloud protocols, it was apparent that in many cases these answers represented words that could have

been guessed in context, whether due to being a cognate, a previous encounter, or a transparent meaning, such as *atop.* Based upon this analysis, it seems that if learners were to be given credit for partial knowledge, then a more sensitive test format that better approximates the natural act of reading and differentiates between guessing types, would provide more accurate raw scores on which to base estimates.

Central to this discussion is the question of which measure of vocabulary knowledge best reflects the natural act of reading; meaning recall or meaning recognition. As Stewart (2014) explains tests of meaning recall record considerably lower scores than meaning recognition tests. Meaning recall tests more closely represent the necessary word knowledge required to read quickly and fluently, as lists of meanings are not provided when encountering a word in a natural context and automaticity is necessary (McLean, Stewart & Batty 2020; Jeon & Yamashita 2014). This is strongly supported by the consensus of the literature (Stoeckel, et.al. 2019; Zhang & Zhang 2020). However, simply providing a word in isolation and asking a test taker to recall a meaning translation in a native language is potentially more difficult to score and inauthentic in design. As words are seldom encountered alone, unless in the context of rote memorisation flashcards, and context is essential to understanding and differentiating vocabulary meanings it is important to use none-defining contexts in meaning recall tests as utilized by McLean Stewart and Batty (2020). Furthermore, the ability to guess or approximate word meanings in context through word parts, similarities, polysemy, cognates, or other factors that significantly reduce a word's learning burden, are important components in learning to read. Therefore, the findings of this study reaffirm the many criticisms of meaning-recognition, multiple-choice formats by showing that the VST facilitates guesses unrelated to knowledge of items. However, it also shows how the VST's, M/C format enables learners to use partial knowledge, or "comprehension vocabulary" (Laufer & Aviad-Levitzky 2017) to 'guess' correct answers. Based upon this evaluation, it seems a test in which items are presented in context and then recalled in a native language may best replicates the natural act of reading. One such example is the vocableveltest.org platform which provides a means of creating automatically scored meaning-recall vocabulary tests which reduce the burden of marking (Mclean et al 2021). Alternatively, computer-administered, serial multiple-choice formats (SMC) in which distractors are shown sequentially and learners are unable to go back to rejected distractors may also provide a more accurate meaning recognition test (Stoeckel, McLean, & Nation, 2020), which can support informed guessing.

Research limitations

Firstly, the homogeneity of the Japanese participants in this study means that it is difficult to generalise results particularly about total guessing quantities to other groups. Also, as participants shared the ability to guess cognates, Japanese loan words, this would have influenced their ability to use some guessing strategies. Other language groups have different quantities of cognates, and research into how these interact with a target language, such as that conducted by Elgort (2013), is important to understanding how these words affect vocabulary size estimates. The second limitation is that although the conclusions of this study are supported by the literature, the total sample size of the self-report tests and think-aloud tests combined (23) was too limited to generalise definitive conclusions on the functioning of the VST in isolation. The researcher would encourage any interested party to develop the self-report phase of the study based upon the following recommendations: 1) Expand the self-report group participant numbers, but keep group size manageable, so that advice can be provided if necessary. 2) Give clear descriptions of the categories, preferably including video examples. Utilising this methodology could provide finer grained and more generalisable data to inform future vocabulary test development.

Conclusion

In this study a think-aloud protocol and self-reports were used to record and identify the types and quantities of guessing occurring in the multiple-choice VST by requiring participants to specify words they did not know and then try to guess the meaning. The findings suggest that allowing guessing in scores increases vocabulary size estimates; this increase in scores varies between test takers but is more problematic at lower scoring levels; and that correct guesses are the result of chance, exploiting the test format, and applying partial knowledge. Seven guessing types were identified, described and evaluated. Three of these; similar words, polysemy and word parts, accounting for approximately 10% of final scores; demonstrated some partial knowledge. Also, in some cases, distractor-triggered responses could have been guessed in a real reading context. However, as the VST has been shown to be inaccurate in both this study, and many others, it would be better to use alternative vocabulary test instruments. New technology can provide the opportunity to effectively administer more sophisticated tests that include partial knowledge such as serial multiple-choice tests (Stoeckel & Sukigara 2018), or streamline the time-consuming marking of highly accurate, contextualized, meaning-recall tests (Mclean et. al. 2021. Moreover, big data can be utilized to create adaptive tests which better target specific populations and purposes. It is time to move on from outdated test models to focus on producing rigorously developed and validated new vocabulary testing instruments (Schmitt, Nation & Kremmel 2020; Stewart et al, 2021)

About the Author

Steven Asquith is a specially appointed associate professor with the Center for Foreign Language Education and Research at Rikkyo University and a doctoral candidate with the University of Illinois at Urbana Champagne. His main research interests are content and language integrated learning (CLIL), progressive pedagogical design, learner autonomy, and vocabulary. ORCID ID: 0000-0002-9154-9247

To Cite this Article

Asquith, S. (2022). An investigation into the roles of guessing and partial knowledge in the vocabulary size test. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 26* (3). https://doi.org/10.55593/ej.26103a15

References

Aitchison, J. (2012). *Words in the Mind: An Introduction to the Mental Lexicon* (4th Edition ed.). Wiley Blackwell.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27* (1), 101-118. https://doi.org/10.1177/0265532209340194

Beglar, D. & Nemoto, T. (2014, November). *Comparing Two Vocabulary Size Tests*. Hand-out on current research presented at the JALT 2014 Conference, Tsukuba, Japan.

Bennett, P., & Stoeckel, T. (2012). Variations in format and willingness to skip items in a multiple-choice vocabulary test. *Vocabulary Education and Research Bulletin*, *1*, 2–3.

Coxhead, A., Nation, P. and Sim, D. (2015). Measuring the Vocabulary Sizes of Native Speakers of English in New Zealand Secondary Schools. *New Zealand Journal of Educational Studies, 50* (1), 121-135. https://doi.org/10.1007/s40841-015-0002-3

Daulton, F. E. (2007). Japanese Learners' built in lexicon of English and its effect on L2 production. *The Language Teacher, 31* (9). https://doi.org/10.21832/9781847690319

EIKEN (2022) The EIKEN Foundation of Japan. https://www.eiken.or.jp/eiken/en/

Elgort, I. (2013). Effects of L1 definitions and cognate status on the Vocabulary Size Test. *Language Testing, 30* (2), 253-272. https://doi.org/10.1177/0265532212459028

Heigham, J., & Croker, R. A. (2009). *Qualitative Research in Applied Linguistics: A Practical Introduction.* Palgrave Macmillan.

Gyllstad, H., McLean, S., & Stewart, J. (2019, July). Empirically investigating the adequacy of item sample sizes of vocabulary levels and vocabulary size tests: A bootstrapping approach. Paper presented at the Vocab@Leuven Conference, Leuven, Belgium.

Gyllstad, H., Vilkaite, L. and Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *International Journal of Applied Linguistics, 166* (2), 278-306. https://doi.org/10.1075/itl.166.2.04gyl

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

Kamimoto, T. (2008). *Nation's Vocabulary Levels Test and its successors: a reappraisal.* Unpublished PhD Thesis, University of Wales: Swansea.

Laufer, B. & Aviad–Levitzky, T. (2017), What Type of Vocabulary Knowledge Predicts Reading Comprehension: Word Meaning Recall or Word Meaning Recognition? *The Modern Language Journal,* 101: 729-741.

Laufer, B., & 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

MacDonald, K. &. Asaba, M. (2015). "I Don't Know' Use and Guessing on the Billingual Japanese Vocabulary Size Test: A Preliminary Report. *Vocabulary Learning and Instruction, 4* (1), 16-25.

McLean, S., Kramer S. and Stewart, J. (2015). An Empirical Investigation of the Effect of Guessing on Vocabulary Size Test Scores. *Vocabulary Learning and Instruction, 4* (1), 16-25. doi: http://dx.doi.org/10.7820/vli.v04.1.mclean.et.al

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. https://doi.org/10.7820/vli.v10.2.mclean

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

Nation, I. S. P. (2013). *Learning Vocabulary in Another Language Second Edition.* Cambridge University Press.

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

Nation, P. &. Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching, 47* (3), 398-403. https://doi.org/10.1017/S0261444814000111

Nation, P. (2012, October 23). *Vocabulary Size Test information and specifications.* Retrieved October 10, 2013 from The Vocabulary Size Test:

https://www.wgtn.ac.nz/lals/resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

Nguyen, L. T. C. & Nation, I. S. P. (2011). A bilingual vocabulary test of English for Vietnamese learners. *RELC Journal, 42* (1), 86-99. https://doi.org/10.1177/0033688210390264

Saldaña, J. (2011). *Fundamentals of qualitative research.* Oxford University Press.

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, 109–120. doi:10.1017/S0261444819000326

Schmitt, N. S., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of the two new versions of the vocabulary levels test. *Language Testing, 18* (1), 55-88. https://doi.org/10.1177/026553220101800103

Stewart, J. (2014). Do Multiple-Choice Options Inflate Estimates of Vocabulary Size on the VST? *Language Assessment Quarterly, 11* (3), 271-282. https://doi.org/10.1080/15434303.2014.922977

Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. (2021). What the research shows about written receptive vocabulary testing - A reply to Webb. Studies in Second Language Acquisition, 43(2), 462-471. doi-10.1017_S0272263121

Stewart, J. &. White, D. (2011). Estimating guessing effects on the Vocabulary Levels Test for differing degrees of word knowledge. *TESOL Quarterly, 45*, 370-380. doi: 10.5054/tq.2011.254523

Stoeckel, T., Bennett, P., & Mclean, S. (2016). Is "I don't know" a viable answer choice on the vocabulary size test? *TESOL Quarterly*, *50*(4), 965–975. https://doi.org/10.1002/tesq.325

Stoeckel, T., McLean, S., & Nation, P. (2020) Limitations of size and levels tests of written receptive vocabulary knowledge. Studies in Second Language Acquisition, 1-23. https://doi.org/10.1017/S027226312000025X

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. https://doi.org/10.1016/j.system.2019.102161.

Stoeckel, T., & Sukigara, T. (2018). A serial multiple-choice format designed to reduce overestimation of meaning-recall knowledge on the Vocabulary Size Test. *TESOL Quarterly,* 52, 1050–1062. doi: 10.1002/tesq.429

TOEIC (2022) ETS. TOEIC. https://www.iibc-global.org/toeic/test/lr.html

Victoria University of Wellington (2022) Paul Nation's resources. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests

Webb, S. (2007). The Effects of Repetition on Vocabulary Knowledge. *Applied Linguistics, 28* (1), 46-65. https://doi.org/10.1093/applin/aml048

Zhang, X. (2013). The I Don't Know Option in the Vocabulary Size Test. *TESOL Quarterly, 47*, 790-811. https://doi.org/10.1002/tesq.98

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

Appendix 1: The modified VST used in the think-aloud tests

Vocabulary Size Test

[Note: The original monolingual 14000, 140 item version of this test is available as a PDF at https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests. This version was adapted to a 70-item version by removing 5 items from each 10-item band and adding a DK option. The test was also re-ordered so that participants would encounter high and low frequency words throughout the test.

Test Instructions

• There are 70 questions in this test.

• Look at the word and an example of the word in use.

• Choose the meaning that most closely matches the highlighted words in the example sentence.

• If you don't know the word do not guess and circle 'no'.

• Once you have circled 'no', please go back and try to guess the correct answer to the best of your ability.

Example:

SHOE: Where is your **shoe**?

a.) the person who looks after you

b.) the thing you keep your money in

c.) the thing you use for writing

d.) the thing you wear on your foot

e.) I don't know

Note: The data below is in graphical form. For readers who require the actual text, it is available here.

1. SEE: They **saw** it.

a) cut
b) waited for
c) looked at
d) started
e) I don't know

2. MAINTAIN: Can they **maintain** it?

a) keep it as it is
b) make it larger
c) get a better one than it
d) get it
e) I don't know

3. SOLDIER: He is a **soldier**.

a) person in a business
b) student
c) person who uses metal
d) person in the army
e) I don't know

4. SCRUB: He is **scrubbing** it.

a) cutting shallow lines into it
b) repairing it
c) rubbing it hard to clean it
d) drawing simple pictures of it
e) I don't know

5. WEEP: He **wept**.

a) finished his course
b) cried
c) died
d) worried
e) I don't know

6. LATTER: I agree with the **latter**.

a) man from the church
b) reason given
c) last one
d) answer
e) I don't know

7. DEFICIT: The company had a large **deficit**.

a) spent a lot more money than it earned
b) went down a lot in value
c) had a plan for its spending that used a lot of money
d) had a lot of money in the bank
e) I don't know

8. DEVIOUS: Your plans are **devious**.

a) tricky
b) well-developed
c) not well thought out
d) more expensive than necessary
e) I don't know

9. EXCRETE: This was **excreted** recently.

a) pushed or sent out
b) made clear
c) discovered by a science experiment
d) put on a list of illegal things
e) I don't know

10. PERIOD: It was a difficult **period.**

a) question
b) time
c) thing to do
d) book
e) I don't know

11. BUTLER: They have a **butler**.

a) man servant
b) machine for cutting up trees
c) private teacher
d) cool dark room under the house
e) I don't know

12. PALETTE: He lost his **palette.**

a) basket for carrying fish
b) wish to eat food
c) young female companion
d) artist's board for mixing paints
e) I don't know

13. WHIM: He had lots of **whims.**

a) old gold coins
b) female horses
c) strange ideas with no motive
d) sore red lumps
e) I don't know

14. BAWDY: It was very **bawdy.**

a) unpredictable
b) enjoyable
c) rushed
d) rude
e) I don't know

15. UBIQUITOUS: Many weeds are **ubiquitous.**

a) are difficult to get rid of
b) have long, strong roots
c) are found in most countries
d) die away in the winter
e) I don't know

16. SOLILOQUY: That was an excellent **soliloquy!**

a) song for six people
b) short clever saying with a deep meaning
c) entertainment using lights and music
d) speech in the theatre by a character who is alone
e) I don't know

17. UPSET: I am **upset.**

a) tired
b) famous
c) rich
d) unhappy
e) I don't know

18. PATIENCE: He has no **patience.**

a) will not wait happily
b) has no free time
c) has no faith
d) does not know what is fair
e) I don't know

19. RESTORE: It has been **restored.**

a) said again
b) given to a different person
c) given a lower price
d) made like new again
e) I don't know

20. JUG: He was holding a **jug.**

a) A container for pouring liquids
b) an informal discussion
c) A soft cap
d) A weapon that explodes
e) I don't know

21. DINOSAUR: The children were pretending to be **dinosaurs.**

a) robbers who work at sea
b) very small creatures with human form but with wings
c) large creatures with wings that breathe fire
d) animals that lived a long time ago
e) I don't know

22. COMPOUND: They made a new **compound.**

a) agreement
b) thing made of two or more parts
c) group of people forming a business
d) guess based on past experience
e) I don't know

23. CANDID: Please be **candid.**

a) be careful
b) show sympathy
c) show fairness to both sides
d) say what you really think
e) I don't know

24. TUMMY: Look at my **tummy.**

a) cloth to cover the head
b) stomach
c) small furry animal
d) thumb
e) I don't know

25. QUIZ: We made a **quiz.**

a) thing to hold arrows
b) serious mistake
c) set of questions
d) box for birds to make nests in
e) I don't know

26. NUN: We saw a **nun**.

a) long thin creature that lives in the earth
b) terrible accident
c) woman following a strict religious life
d) unexplained bright light in the sky
e) I don't know

27. HAUNT: The house is **haunted.**

a) full of ornaments
b) rented
c) empty
d) full of ghosts
e) I don't know

28. COMPOST: We need some **compost**.

a) strong support
b) help to feel better
c) hard stuff made of stones and sand stuck together
d) rotted plant material
e) I don't know

29. PREMIER: The **premier** spoke for an hour

a) person who works in a law court
b) university teacher
c) adventurer
d) head of the government
e) I don't know

30. ACCESSORY: They gave us some **accessories.**

a) papers allowing us to enter a country
b) official orders
c) ideas to choose between
d) extra pieces
e) I don't know

31. THRESHOLD: They raised the **threshold.**

a) flag
b) point or line where something changes
c) roof inside a building
d) cost of borrowing money
e) I don't know

32. OLIVE: We bought **olives.**

a) oily fruit
b) scented pink or red flowers
c) men's clothes for swimming
d) tools for digging up weeds
e) I don't know

33. QUILT: They made a **quilt.**

a) statement about who should get their property when they die
b) firm agreement
c) thick warm cover for a bed
d) feather pen
e) I don't know

34. STEALTH: They did it by **stealth.**

a) spending a large amount of money
b) hurting someone so much that they agreed to their demands
c) moving secretly with extreme care and quietness
d) taking no notice of problems they met
e) I don't know

35. SHUDDER: The boy **shuddered**.

a) spoke with a low voice
b) almost fell
c) shook
d) called out loudly
e) I don't know

36. BRISTLE: The **bristles** are too hard.

a) questions
b) short stiff hairs
c) folding beds
d) bottoms of the shoes
e) I don't know

37. TIME: They have a lot of **time**.

a) money
b) food
c) hours
d) friends
e) I don't know

38. MONOLOGUE: Now he has a **monologue.**

a) single piece of glass to hold over his eye to help him to see better
b) long turn at talking without being interrupted
c) position with all the power
d) picture made by joining letters together in interesting ways
e) I don't know

39. ERRATIC: He was **erratic.**

a) without fault
b) very bad
c) very polite
d) unsteady
e) I don't know

40. NULL: His influence was **null.**

a) had good results
b) was unhelpful
c) had no effect
d) was long-lasting
e) I don't know

41. KINDERGARTEN: This is a good **kindergarten.**

a) activity that allows you to forget your worries
b) place of learning for children too young for school
c) strong, deep bag carried on the back
d) place where you may borrow books
e) I don't know

42. ECLIPSE: There was an **eclipse.**

a) a strong wind
b) a loud noise of something hitting the water
c) The killing of a large number of people
d) The sun hidden by a planet
e) I don't know

43. FIGURE: Is this the right **figure?**

a) answer
b) place
c) time
d) number
e) I don't know

44. HALLMARK: Does it have a **hallmark?**

a) stamp to show when to use it by
b) stamp to show the quality
c) mark to show it is approved by the royal family
d) Mark or stain to prevent copying
e) I don't know

45. PURITAN: He is a **puritan.**

a) person who likes attention
b) person with strict morals
c) person with a moving home
d) person who hates spending money
e) I don't know

46. WEIR: We looked at the **weir.**

a) person who behaves strangely
b) wet, muddy place with water plants
c) old metal musical instrument played by blowing
d) thing built across a river to control the water
e) I don't know

47. AWE: They looked at the mountain with **awe.**

a) worry
b) interest
c) wonder
d) respect
e) I don't know

48. PEASANTRY: He did a lot for the **peasantry**.

a) local people
b) place of worship
c) businessmen's club
d) poor farmers
e) I don't know

49. EGALITARIAN: This organization is **egalitarian.**

a) does not provide much information about itself to the public
b) dislikes change
c) frequently asks a court of law for a judgement
d) treats everyone who works for it as if they are equal
e) I don't know

50. MYSTIQUE: He has lost his mystique.

a) his healthy body
b) the secret way he makes other people think he has special power or skill
c) the woman who has been his lover while he is married to someone else
d) the hair on his top lip
e) I don't know

51. UPBEAT: I'm feeling really **upbeat** about it.

a) upset
b) good
c) hurt
d) confused
e) I don't know

52. MUSSEL: They bought **mussels**.

a) small glass balls for playing a game
b) shellfish
c) large purple fruits
d) pieces of soft paper to keep the clothes clean when eating
e) I don't know

53. YOGA: She has started **yoga**.

a) handwork done by knotting thread
b) a form of exercise for body and mind
c) a game where a cork stuck with feathers is hit between two players
d) a type of dance from eastern countries
e) I don't know

54. COUNTERCLAIM: They made a **counterclaim**.

a) a demand made by one side in a law case to match the other side's dema
b) a request for a shop to take back things with faults
c) An agreement between two companies to exchange work
d) a top cover for a bed
e) I don't know

55. PUMA: They saw a **puma**.

a) small house made of mud bricks
b) tree from hot, dry countries
c) very strong wind that sucks up anything in its path
d) large wild cat
e) I don't know

56. HAZE: We looked through the **haze**.

a) small round window in a ship
b) unclear air
c) strips of wood or plastic to cover a window
d) list of names
e) I don't know

57. SPLEEN: His **spleen** was damaged.

a) knee bone
b) organ found near the stomach
c) pipe taking wastewater from a house
d) respect for himself
e) I don't know

58. REPTILE: She looked at the **reptile**.

a) old hand-written book
b) animal with cold blood and a hard outside
c) person who sells things by knocking on doors
d) picture made by sticking many small pieces of different colors together.
e) I don't know

59. ALUM: This contains **alum**.

a) a poisonous substance from a common plant
b) a soft material made of artificial threads
c) a tobacco powder once put in the nose
d) a chemical compound usually involving aluminium
e) I don't know

60. STONE: He sat on a **stone**.

a) hard thing
b) kind of chair
c) soft thing on the floor
d) part of a tree
e) I don't know

61. TALON: Just look at those **talons!**

a) high points of mountains
b) sharp hooks on the feet of a hunting bird
c) heavy metal coats to protect against weapons
d) people who make fools of themselves without realizing it
e) I don't know

62. ROUBLE: He had a lot of **roubles.**

a) very precious red stones
b) distant members of his family
c) Russian money
d) moral or other difficulties in the mind
e) I don't know

63. JOVIAL: He was very **jovial.**

a) low on the social scale
b) likely to criticize others
c) full of fun
d) friendly
e) I don't know

64. COMMUNIQUE: I saw their **communiqué**

a) critical report about an organization
b) garden owned by many members of a community
c) printed material used for advertising
d) official announcement
e) I don't know

65. POOR: We are **poor.**

a. have no money
b. feel happy
c. are very interested.
d. do not like to work hard
e. I don't know

66. CANONICAL: These are **canonical** examples.

a) examples which break the usual rules
b) examples taken from a religious book
c) regular and widely accepted examples
d) examples discovered very recently
e) I don't know

67. ATOP: He was **atop** the hill.

a) at the bottom of
b) at the top of
c) on this side of
d) on the far side of
e) I don't know

68. MARSUPIAL: It is a **marsupial.**

a) an animal with hard feet
b) a plant that grows for several years
c) a plant with flowers that turn to face the sun
d) an animal with a pocket for babies
e) I don't know

69. AUGUR: It **augured** well.

a) promised good things for the future
b) agreed well with what was expected
c) had a color that looked good with something else
d) rang with a clear, beautiful sound
e) I don't know

70. DRAWER: The **drawer** was empty.

a) sliding box
b) place where cars are kept
c) cupboard to keep things cold
d) animal house
e) I don't know

Appendix 2: Backgrounds, estimated levels, and total scores of think-aloud test subjects

Subject	Background	Level estimate (Formal qualification)	VST Score
A	Airline fight staff, who uses English frequently both inside and outside of work. Often reads books and magazines, and watches TV in English.	Upper intermediate (TOIEC 810)	52
B	Distributions manager at an international trading company. Often uses English at work, particularly email, and enjoys speaking English with his many non-Japanese friends. Lived in the UK for two years.	Upper intermediate (TOIEC 730)	48
C	English teacher, who works at a junior high school. Outside of work enjoys reading a newspaper in English regularly.	Intermediate (NA)	47
D	University student, who recently returned from three months studying English in Canada.	Intermediate (NA)	46
E	English teacher, who works at a junior high school. Although he uses English for work, this is usually not at a level that would increase his vocabulary. English is not used outside of work.	Intermediate (NA)	45
F	English teacher, who works at a junior high school. Outside of work she frequently enjoys watching TV dramas in English.	Intermediate (NA)	44
G	Housewife with a background in international trade. Uses English as the main means of communication at home. Often watches TV in English but seldom reads extensively.	Intermediate (TOIEC 730)	42
H	Fight staff trainer, who enjoys studying English particularly focussing at increasing her TOIEC and EIKEN scores. Does not usually read extensively.	Intermediate (TOIEC (700)	39
I	Office worker, who studies at a weekly conversation class. No formal English schooling beyond compulsory education	Elementary/ Pre int. (NA)	30
J	Housewife, who enjoys studying English at a weekly conversation class. Does not usually practice English outside of class. No formal English schooling beyond compulsory education	Elementary/Pre Int. (NA)	27

*Assessment of levels are estimates based upon 15 years of teaching experience and knowledge of participants, other than Participant H, for an extended period. Test scores were provided by participants.

Appendix 3: Transcripts of the think aloud tests showing examples of the different guessing strategy categories

1. Distractor triggered response	
Test item	62. ROUBLE: He had a lot of **roubles.** a) very precious red stones b) distant members of his family c) Russian money d) moral or other difficulties in the mind e) I don't know
Transcript	**Participant** ro-bles … *hmm kore wakaranai (I don't know this) ruburu (rubles) to chigaou ki ga suru (but it seems wrong)* Russian money (reading the distractor), *ma iiya ruburu ni shiya (what the heck, lets go with ruble)*
Commentary	In this case the participant had difficulty reading and pronouncing the item. Although he knew the Japanese cognate, it was difficult for him to connect it to the English spelling. Upon reading the distractor the participant realised the connection and chose Russian money.

2. Similar words	
Test item	64. COMMUNIQUE: I saw their **communiqué** a) critical report about an organization b) garden owned by many members of a community c) printed material used for advertising d) official announcement e) I don't know
Transcript	**Participant** communiku… communiqué *dake (perhaps)* … (pause while reading).. this one? **Tester** Why? **Participant** communiqué that means communicate so ahh.. its like a word, 'announcement'.. similar to this one
Commentary	In this example the participant based her answer upon the similarity between communicate and communiqué. There is some overlap between this category and distractor triggered responses as in most cases the participant used the distractors to confirm or reject their suspicions.

3. Word part	
Test item	16. SOLILOQUY: That was an excellent **soliloquy!** a) song for six people b) short clever saying with a deep meaning c) entertainment using lights and music d) speech in the theatre by a character who is alone e) I don't know
Transcript	so so soliloquy soli ho… soli means…single… soli hmm, soli *eee toh (sound used when thinking)* song for six people (reading distractor a aloud) *ahh so ka (I see)* ahhh short clever saying with a deep meaning.. oh entertainment using lights and music… speech in theatre.. ah! *kore kana (this one). demo kore wa kore desu (well this is this one)*
Commentary	The participant hesitated reading the item at first but then pronounced it correctly. This may suggest some knowledge of the word, as it is particularly difficult to read with the correct pronunciation. He then correctly identified that soli means single before quietly reading each distractor under his breath. At the point he encountered 'alone' in the distractor he immediately says "ah, this one" and identifies the correct answer.

4. Polysemy	
Test item	ACCESSORY: They gave us some **accessories.** a) papers allowing us to enter a country b) official orders c) ideas to choose between d) extra pieces e) I don't know
Transcript	**Participant** …they gave us some accessories…(long pause) pieces. but.. don't we use accessories for like.. earrings.. (reading the distractors again) papers allowing… *eeeeeeh* I didn't... *hmm ja.. (ok, in that case)* (choosing the correct answer)
Commentary	In this case, the participant was surprised when the definition she expected 'jewelry' was not there. She then had to work back to decide which distractor was closest to the known meaning.

5. Semantic sense	
Test item	26. NUN: We saw a **nun**. a) long thin creature that lives in the earth b) terrible accident c) woman following a strict religious life d) unexplained bright light in the sky e) I don't know

Transcript	***Participant:*** *wakaran kore wa (I don't know).. sozo suru (I'll use my imagination) ...* (long pause) *kore kana, (maybe this one)* **Tester** *eh nande (why?)* **Participant** *nantonaku (a vague idea) mazu ningen daro (firstly I thought it might be a person) religious no kanji ga suru (it has a religious feel)* **Tester** *nanka kita koto aru ka (have you heard it before?)* **Participant** *kono ki ga suru (it has that sense)*
Commentary	The participant first says he doesn't know this item followed by a long pause and a correct guess. When asked why, he replies that he just has a vague sense of the word and may have encountered it before. This vague sense could not reasonably be categorised as knowing a word within a minimum threshold for knowledge as it is unlikely to be effective when reading given the long pause required.

6. Distractor elimination	
Test item	31. THRESHOLD: They raised the **threshold.** a) flag b) point or line where something changes c) roof inside a building d) cost of borrowing money e) I don't know
Transcript	**Participant** threshold *wakaranai ore (I don't know)* flag, point or line where something changes, roof inside a building, cost of borrowing money… flag *ka? roof ka,* point or line between something changes, cost of borrowing money *kore tesuryo ka (Is this commission?)* where something changes *kore kana? (maybe this)* **Tester** *Any reason, or just.. just random?* **Participant** *nanka B to D wa kankei aru kara, kankei chigaou, nittei yo na (b and d are connected, not connected, similar)* flag *de* roof *de wa kore wa* things, *kore wa chotto nan to iiu..* abstract. *(flag and roof are things, this one is, how do you say… abstract)*
Commentary	In this example, the participant uses a process of elimination to reject the concrete things and settle on the correct answer of the 'abstract' notion of a point or line. There is little to no demonstration of actual knowledge of the word, rather using the format of the test to deduct the correct answer.