Reading in a Foreign Language October 2021, Volume 33, No. 2
ISSN 1539-0578 pp. 247–259

Lexical Mastery Thresholds and Lexical Units: A Reply to Laufer

Stuart McLean
Momoyama Gakuin University
Japan

Tim Stoeckel
University of Niigata Prefecture
Japan

In response to McLean (2021), Laufer (2021) makes three claims which are either not supported by research or are based on studies with important limitations. First is that a vocabulary size, instead of a level, can be used to match learners with lexically appropriate materials despite test creators and research not supporting this. Second is that the word family (WF6) is an appropriate definition of the lexical unit if learners know at least 5,000 WF6s. The available evidence suggests that for such learners, knowledge of derivational forms is limited enough that it can result in the incorrect matching of learners to pedagogical materials (McLean, 2018). Additionally, foreign language learners who know 5,000 WF6s are rare. Third is that derivational forms are infrequent enough that knowledge of only a few affixes will support comprehension. This inference results from Laufer and Cobb's (2020) analysis, which has major limitations.

We are sincerely thankful for Laufer's interest in McLean's 2021 publication and for discussing the recent commentary regarding the limitations of levels and size tests (Stewart, et al., 2021; Stoeckel, et al., 2021; Webb, 2021). We hope readers will carefully read all of these works and consider the validity of the arguments based on the evidence presented.

Using Vocabulary Tests for Materials Selection

Laufer (2021) states that "[i]t is easy to set a very high lexical mastery level when a text is composed of 1000 or 2000 most frequent words, and students possess limited vocabularies. It is much more problematic to do so with most text profiles ... and most students' vocabularies … . A total vocabulary size score is a practical alternative" (p. 244). This statement is problematic for several reasons.

First, it contradicts the recommendations of vocabulary levels and size test creators, who state that scores from individual word bands, not total scores, should be used to establish a lexical mastery level or to match learners with materials (Nation, 2016; Schmitt et al., 2001; Webb et al., 2017). For instance, Webb et al. (2017) state "when interpreting scores, it is the scores for the individual levels of the Vocabulary Levels Test (VLT) that are meaningful rather than the scores for all levels combined" (p. 55).

Second, using vocabulary size to match learners with materials is inconsistent with the relevant research. Schmitt et al. (2011) found that if learners know 98% of the tokens within a text, it can usually be comprehended. Accordingly, knowledge of words beyond the 3K level

is irrelevant if the tokens within a text are from the first 3,000 words. A learner's total score on a vocabulary test suggesting that they know 3,000 of the first 5,000 words, does not mean they have mastery of the first three 1,000-word bands (Stoeckel & Bennett, 2015).

Third, Laufer overlooks McLean's (2021) stated context - English as a foreign language (EFL) or expanding circle education settings, where most learners share the same first language (L1), and the generally low-proficiency learners have limited exposure to English. Instead, Laufer concentrates on reading materials that require a vocabulary level of 3,000 WF6s for 95% coverage and 5,000 WF6s for 98% coverage. Table 1 shows that mastery of the 2K level is rare in the settings described by McLean, while mastery of the 3K to 5K levels is extremely rare, even among English and translation majors.

These three points address Laufer's question of whether there is a need for rigid mastery levels for matching texts to learners' levels. Considering research and the recommendations of test creators, if we want to make valid inferences about a learner's lexical mastery level, then there is a need for evidence-supported word-band mastery levels of 95% (to support comprehension), 98% (for meaning-focused input), and 100% (for fluency development).

The pressure for lax mastery thresholds, unsupported by research, might result from two limitations of existing levels tests. First, frequency-based tests include unexpectedly difficult words that are unrepresentative of the level as a whole. For example, the word *fellow* is in the 1K band of the Updated Vocabulary Levels Test (Webb et al., 2017; UVLT). Second, monolingual test formats in which response options are in the second language (L2) can prevent learners from demonstrating knowledge of known words (Stoeckel et al., 2019). Thus, when matching learners with texts, we should consider basing tests and lexical profilers on knowledge- rather than frequency-based lists (Paul Nation, personal communication, August 8, 2021), and when it is not possible to use meaning-recall tests, meaning-recognition tests could offer response options in the L1.

Table 1

Levels of Lexical Mastery

Study	N	Participants	1K	2K	3K	4K	5K
			Students (%) who demonstrated mastery (29/30 or 96.6%) on the UVLT				
Feng & Webb, (2020)	76	English Translation majors in China	48.68%	25.01%	21.05%	*	*
			Mean (*SD*) UVLT score (*k* = 30)				
Jin & Webb, (2021)	140	English Majors in China "relatively advanced users of English within the Chinese EFL context."	29.7 (1.14)	28.4 (1.9)	25.3 (3.4)	19.0 (4.9)	15.0 (5.7)
			Mean (*SD*) VLT score (*k* = 30)				
Boutorwick et al., (2019)	63	EAP learners preparing for undergraduate or graduate education in New Zealand.		24 (3.6)	20.7 (4.4)		15.4 (5.4)
Uchihara & Harada, (2018)	120	Japanese English Medium Instruction learners. Mean TOEIC = 723.9 (*114.5*). Mean TOEFL = 538.8 (*38.0*).		29.1 (0.9)	26.5 (2.6)		18.7 (3.7)

Note. *impossible to calculate

Testing Purpose and Construct Validity

Laufer cites past validation studies to argue that vocabulary size and levels tests are appropriate for a variety of research purposes. While it is true that validation studies have been conducted, we must remember that such studies aim to provide validity of inferences made from a test at a given time, with a given set of learners, and for a specific purpose.

This is sometimes forgotten since written receptive vocabulary tests that are *read* are inappropriately used to establish that *listening* materials are lexically appropriate. This is despite (a) evidence that learners' written and spoken receptive vocabulary knowledge differ significantly (Mizumoto & Shimamoto, 2008) and (b) explicit caution like that from Beglar (2010) that "using the test to measure test-takers' listening vocabulary size is not recommended as reading and listening vocabulary sizes can vary considerably" (p. 114). Additionally, whereas some validation studies have indeed established strong psychometric test characteristics, others have found a tenuous relationship between test scores and the lexical knowledge that such scores are supposed to represent (e.g., Kremmel & Schmitt, 2016).

Regarding levels tests, no study has validated the mastery thresholds suggested by test creators in terms of real-world language use. This would be, for instance, that learners who achieve mastery scores on the first two levels of a levels test possess sufficient vocabulary knowledge to comprehend materials written at the 1K and 2K levels. There are several reasons why existing tests may not make this distinction for many expanding circle EFL learners. First, existing tests use meaning-recognition (multiple-choice or matching) formats, which overestimate the type of vocabulary knowledge needed for reading and listening (McLean et al., 2014; McLean et al., 2015; McLean et al., 2020; Stewart et al., 2017; Zhang & Zhang, 2020). Second, the existing tests represent 1,000 words with too few items, reducing the accuracy of scores relative to learners' true knowledge of the target bands (Stoeckel et al., 2021). Third, 1,000-word bands are too coarse for most EFL learners (even some English majors, Table 1), especially when measuring knowledge of high-frequency words (Kremmel, 2016). Fourth, WF6-based tests overestimate learners' knowledge by assuming the understanding of derivational forms (e.g., *useless*, *useful*, *usable*) from knowledge of the baseword (e.g., *use*; Stoeckel et al., 2021). Finally, WF6-based profilers underestimate text difficulty by grouping less frequent, commonly unknown derivational forms (e.g., *useful*, *usable*) within high-frequency families (e.g., *use*). Limitations of WF6 are further discussed below.

Lexical Units

Problems with overestimating derivational knowledge

When matching EFL learners with lexically appropriate materials, the overestimation of derivational knowledge is a greater issue than its underestimation for at least two reasons. First, Schmitt et al.'s (2011) findings that a 1% decline in coverage reduces comprehension by 2.3% suggest that a small overestimation of word knowledge has a disproportionate impact on comprehension. Second, the coverage window for unassisted comprehension is only 5%—between 95% and 100% coverage. Thus, any overestimation of word knowledge can quickly result in unintended reading difficulties. In contrast, underestimation of vocabulary knowledge will result in greater coverage, something which is still beneficial for L2 reading development.

As previously mentioned, when knowledge of WF6 derivational forms is assumed in vocabulary tests and lexical profilers, unknown derivational forms cause an overestimation of the percentage of known tokens in a text. Learners in EFL settings have been observed to comprehend around only 60% of derivational forms containing frequent affixes when the baseword or another family member is known (Table 2). Additionally, coverage figures usually assume that all proper nouns and homoforms are known, but research suggests that this is not always the case (Brown, 2013; Klassen, 2021).

In research, if the 95%, 98%, and 100% figures for the lexical load of a text for language-focused learning, meaning-focused input, and fluency development, respectively, are referred to, then these figures should be accurately operationalised.

Table 2

Written Receptive Meaning-recall Knowledge of Derivational Forms Featuring Frequent Affixes (percent correct); after Stoeckel, McLean, and Nation (2020)

	Ward & Chuenjundaeng (2009)		McLean (2018)			
Affix	Participants		Participants			
	Low-group	High-group	All	Beginner	Intermediate	Advanced
-ly***						
-ion***	58.5	31.1				
-er***	66.9	94.2				
-y***						
-al***			84.5	79.8	86.4	88.2
re-***			79.7	66.5	83.7	98.8
un-**						
-age**			22.7	9.5	25.0	64.7
-ness**						
-ity***	41.2	57.5				
-ate*						
-in*						
-ant*						
Mean	55.5	60.9	62.3	52.0	65.0	83.9

Note. Affixes identified as being among the ten most common affixes of English by *Sánchez-Gutiérrez et al. (2018), **Laufer and Cobb (2020), and by ***both sets of authors.

Research on learners' knowledge of derivational forms

Eight studies have investigated L2 English learners' knowledge of derivational forms, all finding incomplete understanding of such forms, including those containing the most common affixes (Table 2). This body of research suggests that WF6 is inappropriate among EFL and expanding circle settings, and as a general lexical unit. Laufer et al. (2021) found a significant ($p < .01$, $d = 0.69$) difference between knowledge of basewords and derivational forms among B1 level learners under the Common European Framework of Reference, but not among B2 learners, suggesting limited derivational knowledge may be related to proficiency. However, the study has several unreported limitations. First, 13 of 60

derivational form test items erroneously employed basewords or inflectional forms instead of derivational forms (Figure 1). Laufer et al. (2021) referred to identical baseword forms that tested a different part of speech (POS) as "'derived' words with 0 affix" (p. 11). However, this is not in line with Bauer and Nation's (1993) treatment of baseword forms of various POS, nor other research by Laufer, for example, Laufer and Cobb (2020). Second, a testing effect was facilitated because the easier baseword test came before the more difficult derivational form test and because the two tests often used similar distractors (Figure 1). Third, the multiple-choice format sometimes enabled learners to answer correctly by using baseword knowledge without knowing the derivational form (Figure 2). Fourth, the derivational forms test sometimes did not assess knowledge of the morphology of derivational forms (e.g., *-able*) because most or all answer options contained the meaning of the target word's affix (e.g., *can*) rather than other affix meanings to act as distractors (Figure 3).

Figure 1

UPSET: I am **upset**.

a. tired
b. famous
c. rich
d. unhappy

UPSET: Will it **upset** me?

a. make me tired
b. make me famous
c. make me rich
d. make me unhappy

Note. Baseword (left) and derivational (right) items*;* The derivational form item erroneously assesses the baseword, and the wording of the response options is similar between items.

Figure 2

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

SHOELESS: He was **shoeless**.

a. without someone to look after him
b. with something to keep his money in
c. with the thing you use for writing
d. without anything on his feet

Note. In the derivation item (right) learners could find *shoe* within *shoeless* and then select the correct answer because of the presence of the *feet* without knowing the meaning of *shoeless*

Figure 3

DRIVE: He **drives** fast.

a. swims
b. learns
c. throws balls
d. uses a car

DRIVABLE: It is **drivable**.

a. can swim
b. can be learned
c. can be thrown like a ball
d. can be used by a car

Note. Paired baseword (left) and derivational (right) items. The derivational item does not assess understanding of the affix *-able* because it is reflected in all four response options.

The importance of research setting and instruments to the validity of inferences

Perhaps Laufer's conclusion regarding the appropriateness of WF6 differs from that of many authors who have investigated L2 English learners' knowledge of derivational forms because of Laufer's over-optimistic view of EFL and expanding circle learners. Laufer refers to learners with mastery of the first 3,000 words as *low level*. However, as previously discussed (Table 1), this level of mastery is rare among EFL learners. Laufer continues, "learners do possess morphological knowledge which improves and becomes quite good when they reach 5000 word knowledge" (p. 245). Perhaps, but in the setting described by McLean (2021), this is extremely unusual.

Laufer criticises the quality of some research, stating that "[u]nderstanding of derived words in texts is not reflected by tests of stemless affixes, or of infrequent derived words presented in isolation and in clueless sentences" (p. 244). While we agree that more robust research is necessary, it seems prudent to select a lexical unit based on available research and the teaching or research goals.

Laufer's view is that including infrequent WF6 derivational forms is a limitation of studies examining knowledge of WF6 members. Though we believe this is misconceived, filtering data so that it includes learner knowledge of derivational forms containing only high-frequency affixes still does not support the use of WF6. Using WF6 assumes that learners who can comprehend the baseword or other WF6 members, can receptively infer the meaning of all WF6 constituents with little or no effort (Bauer & Nation, 1993), *regardless of the frequency of the derivational form* (Paul Nation, personal communication, March 22, 2021). Thus, research that considers the validity of WF6 is concerned with learners' ability to comprehend basewords and their associated derivational forms, and *the frequency of the derivational forms is irrelevant* (Paul Nation, personal communication, March 22, 2021). If derivational form frequency significantly influences comprehension, it is evidence that derivational forms are learned and understood as whole words and not through applying affix knowledge to known basewords. This would, in fact, be evidence against the WF6 construct as conceived by Bauer and Nation.

Regarding item format, we agree with Laufer that tests of stemless affixes (e.g., the form section of Sasao & Webb's 2017 Word Parts Level Test) do not reflect the ability to comprehend derived forms when reading. We also acknowledge limitations to research that has not presented target words in at least somewhat natural reading contexts. Until the efficacy of such contextualized tests has been established, however, we advocate the use of existing meaning-recall (L2 to L1 translation) formats. This is because, as stated by Laufer herself, "[t]he recall of meaning... resembles word interpretation in reading, where the meaning of words in a text need to be recalled" (Aviad-Levitzky, et al., 2019, p. 353), a view that is supported by research (McLean, et al., 2020).

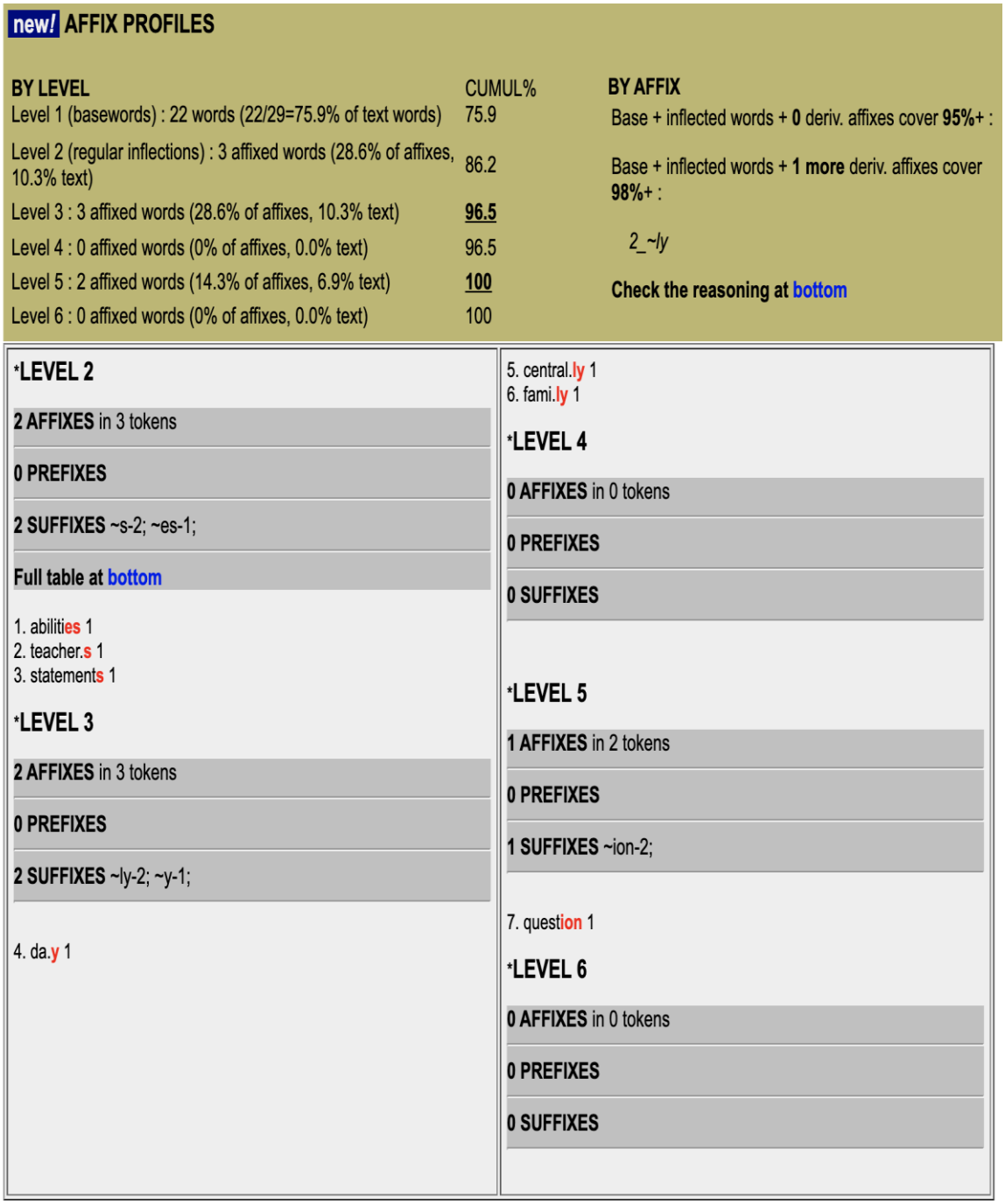
Limitations of Laufer and Cobb's study

Laufer cites Laufer and Cobb (2020) to argue that limited derivational knowledge is of little concern as (a) derivational forms are rare, and (b) a few commonly known affixes (i.e., *-ly, -ion, -er*) provide most of the coverage from derivational forms. However, there are reasons to treat this study with caution. First, it used a corpus of just 25 texts (243,731 tokens), limiting its representativeness. Second, Morpholex, the tool used in the study to analyze texts, was and remains of limited accuracy. For instance, Morpholex sometimes counts *day* and

question as derivational forms (Figure 4). Additionally, Morpholex commonly treats derivational forms that end in inflectional affixes as inflectional and not derivational forms. For example, *abilities* (a derivation of *able*), *teachers* (a derivation of *teach*), and *statements* (a derivation of *state*) are all counted as inflectional forms (Figure 4). This problem affects the classification of many words: we found that among members of the first 1,000 WF6s, 639 derivational forms are wrongly classified as inflectional. Similarly, Morpholex occasionally fails to count affixes within derivational forms containing multiple affixes (e.g., *centrally*, composed of *center* + *al* + *ly*, Figure 4).

Figure 4

Output from Morpholex Showing Misclassifications



How these limitations of Morpholex affected the results of Laufer and Cobb's analysis is unclear, but to better understand the problem, we reanalyzed one text from their corpus (the 2010 paper by Laufer & Ravenhorst-Kalovski), manually coding each token according to (a) the affixes it contains and (b) its classification as a baseword, inflection, or derivational form. Figure 5 shows the most frequent derivational affixes in this text according to Laufer and Cobb and to our counts. While some counts are similar (e.g., *-age*, 121 versus 123 tokens), others are wildly different (e.g., *-er*, 16 to 103). Laufer and Cobb did not report the individual words in which each affix appeared, but for transparency, we provide such a list for four affixes (Table 3; space limitations prevent a comprehensive listing.) From this, some of Morpholex's limitations become apparent. For instance, the presence of *comprehension* under *-ion* appears to explain the difference in counts for that affix. Apparently, when Morpholex strips *-ion* from the word, it does not recognize *comprehens* as a free morpheme, and so *comprehension* is classified as a baseword. This is contrary to both Bauer and Nation's scheme, where base allomorphy is permitted with the *-ion* affix, and to how *comprehension* is listed in the BNC/COCA lists used in Laufer and Cobb's study. As another example, the huge difference in items classified as having *-er* might be because Morpholex is unable to 'find' the affix when it is followed by a plural *-s* (e.g., *teachers*, Figure 4).

Figure 5

Token Counts for the Most Frequent Derivational Affixes in Laufer & Ravenhorst-Kalovski (2010)

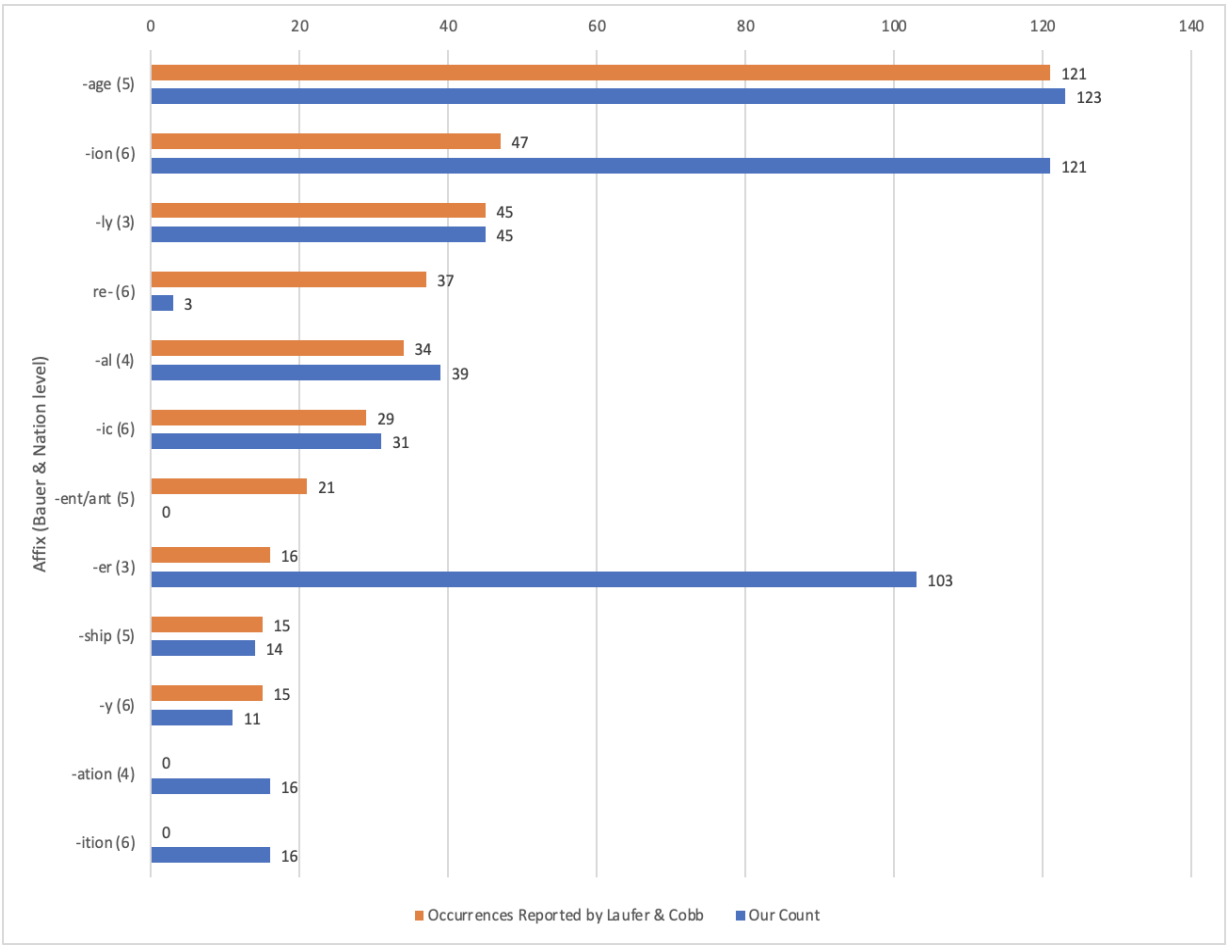


Table 3

Tokens with -ion, -er, -ation, and -ition in Laufer & Ravenhorst-Kalovski (2010).

-ion (6)	n	-er (3)	n	-ation (4)	n	-ition (6)	n
COMPREHENSION	76	LEARNERS	64	INFORMATION	5	ADDITIONAL	14
RELATIONSHIP	14	LEARNER	8	IMPLICATIONALLY	2	DEFINITION	1
REGRESSION	4	TEACHERS	5	REGISTRATION	1	ADDITION	1
EXEMPTION	4	READERS	5	RECATEGORIZATION	1		
TRANSLATION	2	READER	5	OVERSIMPLIFICATION	1		
SUGGESTIONS	2	TAKERS	3	MISREPRESENTATION	1		
IMPLICATIONS	2	RESEARCHERS	3	EXTRAPOLATION	1		
EDUCATIONAL	2	DESIGNERS	3	EXPLANATIONS	1		
EDUCATION	2	PROFILER	2	DEVIATION	1		
DISCUSSION	2	WRITERS	1				
CORRELATION	2	SPEAKERS	1				
SELECTION	1	REVIEWER	1				
RELATION	1	PREDICTOR*	1				
INVESTIGATION	1	EDUCATORS*	1				
INTERACTION	1						
INNOVATIONS	1						
EVALUATION	1						
CONNECTION	1						
CALCULATION	1						
APPROXIMATION	1						
INFLECTIONS	1						
DERIVATIONS	1						
TOTAL	123		103		14		16

Note. The numbers in parentheses indicate the Bauer and Nation (1993) affix level.
* Bauer & Nation do not discuss *-or* or *-ar* as variants of the -er affix. However, the BNC/COCA lists include words employing this affix, so we have included them here.

Third is the unusual way in which Laufer and Cobb estimated the number of derivational affixes a learner would need to know to reach 95 or 98% text coverage. They assumed that knowledge of only *the most frequently occurring derivational affixes* in a given text is sufficient as an estimate of the *total number of affixes* a learner needs to know to understand the text. If such an approach were used to estimate the *lexical* knowledge needed for text comprehension, we might claim that knowing just 1,000 words is sufficient for understanding an academic journal article because fewer than 1,000 unique families are used in the text. A more appropriate method would be to compare the profile of affixes present in a text to a frequency- or knowledge-based list of all affixes, and assume that understanding of all affixes in the list up to the point at which 95 or 98% coverage is reached would be needed for comprehension. Such an approach would be consistent with how we estimate the vocabulary knowledge needed for comprehension.

Fourth, Laufer and Cobb's estimates of the affixational knowledge needed to reach critical coverage percentages are based on the assumptions that (a) learners know 100% of the basewords in a text and (b) knowledge of derivational affixes guarantees understanding of

derivational forms which comprise those affixes plus known basewords. Neither of these assumptions are supported by evidence (McLean, 2018).

Further research is certainly merited, but there are clear reasons to doubt Laufer and Cobb's (2020) arguments that derivational forms are rare, and that a few commonly known affixes account for most of the coverage from derivational forms.

Conclusion

We have discussed the appropriateness of (a) size versus levels tests for matching learners with materials and (b) use of WF6 as a definition of the lexical unit with EFL learners. We have also questioned research supporting the notion that derivational forms are unproblematic for learners because such forms are infrequent and employ mostly common affixes. We hope this discussion can be of some value in moving the field toward more clarity in how we operationalize constructs in research and in more precise measurement.

References

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new computer adaptive test of size and strength (CATSS): Development and validation. *Language Assessment Quarterly, 16*(3), 345–368. https://doi.org/10.1080/15434303.2019.1649409

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279 https://doi.org/10.1093/ijl/6.4.253

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

Boutorwick, T. J., Macalister, J., & Elgort, I. (2019). Two approaches to extensive reading and their effects on L2 vocabulary development. *Reading in a Foreign Language*, *31*, 150–172. https://nflrc.hawaii.edu/rfl/item/423

Brown, D. (2013). Types of words identified as unknown by L2 learners when reading. *System*, *41*(4), 1043–1055. https://doi.org/10.1016/j.system.2013.10.013

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition, 42*(3), 499–523. https://doi.org/10.1017/S0272263119000494

Jin, Z., & Webb, S. (2020). Incidental vocabulary learning through listening to teacher talk. *The Modern Language Journal*, *104*(3), 550–566. https://doi.org/10.1111/modl.12661

Klassen, K. (2021). Proper name theory and implications for second language reading. *Language Teaching*, 1–7. https://doi.org/10.1017/S026144482100015X

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. (2021). Lexical thresholds and alleged threats to validity: A storm in a teacup? *Reading in a Foreign Language. 33*, 238–246.

Laufer, B., & Cobb, T. (2020). How much knowledge of derived words is needed for reading? *Applied Linguistics, 41*(6), 971–998. https://doi.org/10.1093/applin/amz051

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*, 15–30.

Laufer, B., Webb, S., Kim, S. K., & Yohanan, B. (2021). How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *ITL-International Journal of Applied Linguistics, 17*(2), 229–258. https://doi.org/10.1075/itl.20020.lau

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction, 3*(2), 47–55. doi: 10.7820/vli.v03.2.mclean.et.al

McLean, S., Kramer, B. & Stewart, j. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction, 4*(1), 1–10. http://dx.doi.org/10.7820/vli.v04.1.mclean.et.al.

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics, 39*(6), 823–845. https://doi.org/10.1093/applin/amw050

McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalized. *Reading in a Foreign Language, 33*, 126–140. https://nflrc.hawaii.edu/rfl/item/528

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

Mizumoto, A., & Shimamoto, T. (2008). A comparison of aural and written vocabulary size of Japanese EFL university learners. *Language Education & Technology, 45*, 35–51. https://doi.org/10.24539/let.45.0_35

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods, 50*(4), 1568–1580. https://doi.org/10.3758/s13428-017-0981-8

Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research, 21*(1), 12–30. https://doi.org/10.1177/1362168815586083

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Stewart, J., McLean, S., & Kramer, B. (2017). A response to Holster and Lake regarding guessing and the Rasch Model. *Language Assessment Quarterly, 14*(1), 69–74. doi-10.1080/15434303.2016.1262377

Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. G. (2021). What the research shows about written receptive vocabulary testing: A reply to Webb. *Studies in Second Language Acquisition, 43*(2), 462–471. https://doi.org/10.1017/S0272263121000437

Stoeckel, T., & Bennett, P. (2015). A test of the New General Service List. *Vocabulary Learning and Instruction, 4*(1), 1–8. http://vli-journal.org/wp/wp-content/uploads/2021/08/VLI_4_1_1_stoeckel_bennett.pdf

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition, 43*(1), 181–203. 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, 87*, 102–161. https://doi.org/10.1016/j.system.2019.102161

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge Acquisition and applications. *System, 37*(3), 461–469. https://doi.org/10.1016/j.system.2009.01.004

Webb, S. (2021). A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition, 43*(2), 454–461. https://doi.org/10.1017/S0272263121000449

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

Uchihara, T., & Harada, T. (2018). Roles of vocabulary knowledge for success in English medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly, 52*(3), 564–587. https://doi.org/10.1002/tesq.453

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

Acknowledgements

The authors would like to thank Jeff Stewart, Dale Brown, and Phil Bennett for their feedback on this article. We would also like to thank Dale Brown for his assistance in analyzing the Laufer & Ravenhorst-Kalovski (2010) text. Any remaining faults are the sole responsibility of the authors. This research was supported by two Grants-in-aid for Scientific Research (20K00792 and 20K00898) from the Japan Society for the Promotion of Science. Finally, we would like to thank Batia Laufer for sharing the instruments, and for providing details on the order that instruments were given in Laufer, Webb, Kim, & Yohanan (2021).

About the Authors

Stuart McLean is interested in vocabulary, reading, and listening research, and the importance of construct validity within research design. He is currently making online self-marking form-recall and meaning-recall (orthographic and phonological) vocabulary levels tests, that allow teachers to create levels tests based on various (a) lists, (b) word-band sizes, (c) band ranges, and (d) sampling ratios. Teachers can download automatically marked responses, actually typed responses, and the time taken to complete responses. Presently tests are designed for Japanese learners studying English, and English speakers learning Spanish (vocableveltest.org).
E-mail: stumc93@gmail.com

Tim Stoeckel's main research interests include vocabulary learning and testing, word list development, reading, and language fluency. E-mail: stoeckel@unii.ac.jp