
Assessing Vocabulary on the Duolingo English Test



Duolingo Research Report DRR-24-01
Original publication: April 19, 2024 (16 pages)
Current publication: May 16, 2025 (15 pages)
englishtest.duolingo.com/research

Yena Park*, Ramsey Cardwell*, Ben Naismith*

Abstract

This paper describes how the Duolingo English Test (DET) assesses vocabulary knowledge, a key component for evaluating English language proficiency. Utilizing the extended Evidence-Centered Design framework (e-ECD; Arieli-Attali et al., 2019) within the DET's theoretical assessment ecosystem (Burstein et al., 2022), the DET ensures a comprehensive approach to vocabulary task design, administration, and scoring. Here we examine two specific task types—*Read and Select* and *Fill in the Blanks*—designed to assess the breadth, depth, and fluency of test takers' vocabulary knowledge, focusing on specific dimensions of productive and receptive vocabulary knowledge. Together, this combination of vocabulary task types ensures relevant construct coverage, contributing to the overall validity, reliability, and efficiency of the DET's vocabulary assessment.

Keywords

Duolingo English Test, testing vocabulary, vocabulary assessment, vocabulary knowledge

*Duolingo, Inc.

Corresponding author:

Yena Park
Duolingo, Inc. 5900 Penn Ave
Pittsburgh, PA 15206, USA
Email: englishtest-research@duolingo.com

Contents

1	Introduction	3
2	The e-Proficiency Model	3
2.1	Theoretical Descriptions	4
2.2	Vocabulary in the CEFR	5
3	The e-Task Model	6
3.1	Read and Select	7
3.2	Fill in the Blanks	8
3.3	Vocabulary in Other Task Types	9
4	The e-Evidence Model	10
4.1	Grading and Scoring	11
4.2	Relationship to the KSAs	11
5	Conclusion	11

1 Introduction

The Duolingo English Test (DET) is a digital-first, computer-adaptive, high-stakes English proficiency test, widely used for admission to English-medium universities. The DET currently employs 14 different task types to assess English proficiency for both general and academic contexts (Cardwell et al., 2024). Effectiveness and efficiency are prioritized at all stages of the DET’s design process, from task development to administration to scoring.

The focus of this white paper is to describe how the DET assesses the construct of vocabulary knowledge.* While multiple DET task types require test takers to leverage vocabulary knowledge, we center our description around two task types that target vocabulary knowledge as the primary construct: *Read and Select* and *Fill in the Blanks*. *Read and Select* assesses the breadth and fluency of receptive vocabulary knowledge requiring test takers to decide in a short period of time whether the presented string of letters constitutes a real English word (e.g., “constitute”) or a fake word (e.g., “flarpulous”). *Fill in the Blanks* assesses both the breadth and depth of vocabulary knowledge, along with fluency, by requiring test takers to input missing letters to produce a word that fits the context of a sentence. In addition to these two vocabulary-focused task types, vocabulary is also a secondary construct of other task types including *Read and Complete* (aka C-test), *Interactive Reading* (Park et al., 2022), writing tasks (Goodwin et al., 2023), and speaking tasks (Park et al., 2023). These task types collectively assess different aspects of test takers’ receptive and productive command of vocabulary in addition to other language skills.

The description of these vocabulary-focused task types is framed within an e-ECD framework (Arieli-Attali et al., 2019; Langenfeld et al., 2022). The e-ECD framework constitutes part of the DET’s assessment ecosystem (Burstein et al., 2022) as part of the larger measurement framework. Under this framework we discuss activities and evidence supporting three models:

1. the e-Proficiency model, which describes the intended knowledge, skills, and abilities (KSAs) that the test measures
2. the e-Task model, which describes the design features of the task that are intended to elicit performance related to the KSAs as described by the e-Proficiency model
3. the e-Evidence model, which describes the statistical evidence linking task performance to the relevant KSAs of the e-Proficiency model

2 The e-Proficiency Model

In this section, we describe the e-Proficiency model for vocabulary knowledge on the DET. The e-Proficiency model specifies the Knowledge, Skills, and Ability (KSA), which are defined within the e-ECD framework as “latent capacities” targeted by these tasks (Arieli-Attali et al., 2019, p. 2). For the DET, the e-Proficiency model specifies the vocabulary construct based primarily on Nation’s (2001, 2013, 2022) model of form, meaning, and use, as well as the

*Here we use the term vocabulary in the broad sense, also referred to as lexis, which includes the knowledge and use of words, multi-word units, and the connections between them.

relevant scales and descriptors from the Common European Framework of Reference (CEFR; Council of Europe, 2001, 2020).

Vocabulary knowledge is integral to language proficiency and communicative competence. Numerous studies have confirmed the essential role of vocabulary knowledge for other skills in English including reading (e.g., Moghadam et al., 2012), listening (e.g., Bonk, 2000), speaking (e.g., Milton, 2010), and writing (e.g., Dabbagh & Enayat, 2019). Scores on vocabulary size tests have also been found to be significant predictors of university grade point average (GPA), both for second-language (L2) English speakers studying in an anglophone country (Masrai & Milton, 2018; Milton & Treffers-Daller, 2013) and for students at an English-medium university in a non-anglophone country (Harrington & Roche, 2014).

2.1 Theoretical Descriptions

Vocabulary knowledge is a multifaceted construct (Clenton & Booth, 2020; Daller et al., 2007; Qian & Lin, 2020). Two common ways to conceptualize vocabulary knowledge are along the dimensions of breadth versus depth and receptive versus productive.

Breadth was traditionally defined as “the number of words for which the person knows at least some of the significant aspects of meaning” (Anderson & Freebody, 1981, p. 93). In other words, breadth represents the number of words a person has any knowledge of. In contrast, depth refers to how well each word is known in terms of different aspects (to be described further). In addition to breadth and depth, fluency (aka access) has more recently been included as a dimension in this model (Daller et al., 2007); fluency in this context refers to how quickly vocabulary items can be accessed. Productive knowledge refers to the aspects of word knowledge that facilitate a word’s effective use in speech or writing, whereas receptive knowledge refers to the aspects facilitating the comprehension of a word one hears or reads (Laufer et al., 2004).

Nation’s (2001, 2013, 2022) model of vocabulary knowledge captures both the breadth/depth distinction and the receptive/productive distinction (see Table 1). What it means to know a word consists of knowledge in three categories: form, meaning, and use. These three categories are broken down further into nine aspects: form and meaning, concept and referents, associations (meaning); spoken, written, and word parts (form); and grammatical functions, collocations, constraints on use (use). The *form and meaning* aspect (i.e., the form-meaning link; knowing what the word means) in the *Meaning* category is the aspect most closely associated with breadth, whereas the remaining aspects in the model constitute depth (Read & Dang, 2022), including associations, forms (as it pertains to orthography, pronunciation, and morphology), collocations, grammatical functions, and concepts and referents (e.g., polysemy; González-Fernández & Schmitt, 2019; Schmitt, 2014). Each of these aspects has a receptive layer and a productive layer. For instance, the receptive layer of the *form and meaning* aspect in the *Meaning* category involves encountering a word and knowing what it means, while the productive layer could involve seeing the meaning (for example a definition) and being able to recall the word to express that meaning.

Table 1. Reproduction of Nation’s (2022) model of form, meaning, and use with examples

Category	Aspect	Mode	Description
Form	Spoken	R	Recognizing the word form when it is heard
		P	Saying the word with correct pronunciation (both segmental and suprasegmental aspects of pronunciation)
	Written	R	Recognizing the word in its written form when reading
		P	Writing the word with correct spelling
	Word parts	R	Recognizing affixes (e.g., <i>un-</i> , <i>-ed</i>) and the base of a word and knowing how they relate to the meaning
		P	Constructing the word using the right affixes (e.g., forming the antonym of <i>appropriate</i> using the prefix <i>in-</i> , not <i>un-</i> or <i>dis-</i>)
Meaning	Form and meaning	R	Knowing what the word means
		P	Producing the word to express the meaning
	Concepts and referents	R	Knowing the multiple meanings of the word
		P	Producing the word in different contexts to express the range of the word’s meanings
	Associations	R	Knowing other words related to the word like antonyms and synonyms
		P	Producing synonyms and antonyms for the word
Use	Grammatical functions	R	Recognizing that a word has been used correctly
		P	Using the word correctly in a sentence
	Collocations	R	Recognizing other words that occur commonly with the word
		P	Producing words that commonly occur with the word
	Constraints on use	R	Recognizing the nuances of the word
		P	Making a decision on whether to use the word or not given the situation

2.2 Vocabulary in the CEFR

In order to maximize coverage of language constructs, development of the DET is also based on relevant descriptors from the CEFR (Council of Europe, 2001, 2020). The most

directly relevant scales for vocabulary knowledge in the CEFR scale are *Vocabulary range* and *Vocabulary control*, both of which fall under the heading of linguistic competences. The CEFR representation of vocabulary is largely consistent with that of Nation's model in that both explicitly reflect the concepts of breadth and depth (related to *range* and *control* in the CEFR; Milton, 2013) and production and reception. However, the CEFR vocabulary scales complement Nation's (2022) model by focusing on the contexts and tasks comprising actual vocabulary use as opposed to Nation's focus on latent vocabulary knowledge.

As an example, the *Vocabulary range* C1 descriptor is "Can select from several vocabulary options in almost all situations by exploiting synonyms of even words/signs less commonly encountered", and the corresponding C1 control descriptor is "Occasional minor slips, but no significant vocabulary errors" (Council of Europe, 2020, p. 131). Other scales that are less explicitly vocabulary-focused but clearly tied to vocabulary proficiency include *General linguistic range*, *Orthographic control*, *Overall written production*, *Overall oral comprehension*, *Overall reading comprehension*, *Grammatical accuracy*, and *Propositional precision*. The CEFR does not explicitly assign vocabulary items to CEFR levels. However, such a mapping was undertaken by the [English Vocabulary Profile project](#), using learner corpus data to find at which CEFR level learners start to use specific words and phrases.

The CEFR has several primary implications for the e-Proficiency model of vocabulary. First, a speaker's vocabulary repertoire increases over the proficiency levels, with lower proficiency (CEFR A1/A2) characterized by a small repertoire focused on common, familiar, and concrete concepts. Increasing proficiency corresponds to an enlargement of one's vocabulary repertoire, incorporating vocabulary for more specialized and abstract concepts. Higher proficiency is also characterized by increasingly consistent appropriacy and fluency of words from the expanding vocabulary repertoire.

3 The e-Task Model

The e-Task model specifies the features of tasks that are presumed to elicit observations that allow inferences about the target KSAs (Arieli-Attali et al., 2019). We use Nation's (2001, 2013, 2022) model of form, meaning, and use, along with the CEFR scales of vocabulary range and control, to design tasks that elicit different dimensions of vocabulary knowledge.

Vocabulary knowledge is a key component of the broad construct of academic English proficiency that the DET assesses and, as in the CEFR, plays a role in all subconstructs of specific language skills (e.g., reading). As such, assessment of vocabulary knowledge is not limited to a single task type within the DET, but is embedded across multiple task types. This section will focus on two task types that target vocabulary knowledge as their primary construct, detailing their task specifications and supporting evidence from the research literature. The section concludes with a brief description of the role of vocabulary knowledge in other DET task types.

3.1 Read and Select

Read and Select[†] shows a word to test takers who must respond either “Yes” if the word is a real English word or “No” if the word is not a real English word (Figure 1). Test takers are shown up to 18 words in a row and are expected to respond to each item within five seconds.

⌚ 0:05 for this question

Is this a real English word?

somether

 Yes	 No
--	---

Figure 1. A “Read and Select” item

Read and Select features a mix of real and fake English words to address the issue of overestimating proficiency (Beeckmans et al., 2001; Meara, 1992; Meara & Buxton, 1987; Read, 2007). The bank of real words consists of words sampled from a large reference corpus, including general words (typical in registers outside of academic English) and academic words (frequent and dispersed across a range of academic disciplines; Egbert et al., 2020; Gardner & Davies, 2014). Fake words are generated through recurrent neural networks that learn what real words look like and create new fake ones (Settles et al., 2020). Item parameters are estimated for all words based on a featurized three parameter logistic (3PL) model with fixed chance parameter using response data. Some of the features in the 3PL model include those pertaining to word frequency (Milton, 2009), information related to CEFR levels, and neighborhood density (Hashimoto & Egbert, 2019), all of which have been shown to have sizeable contributions to word difficulty. Each (pseudo-)word is administered adaptively, meaning that performance on a word informs the difficulty of subsequent words.

Read and Select employs the format of a Yes/No test (Meara & Buxton, 1987) and measures the breadth of receptive vocabulary knowledge (Beeckmans et al., 2001), corresponding to the CEFR scale of vocabulary range. Additionally, the time limit of five seconds per word on *Read and Select* indicates that the task taps into the dimension of lexical fluency (Daller et al.,

[†] As of April 2, 2024, *Read and Select* has been revised to show only a single word at a time, instead of grouping multiple words together.

2007). In particular, within Nation's (2022) model, the task assesses knowledge of spelling and word parts from the *Form* category and grammatical functions (i.e., the pattern in which the word occurs) from the *Use* category. As we would expect, performance on the Yes/No test has been shown to correlate moderately with other measures of vocabulary knowledge that tap into form and meaning, such as multiple-choice tasks (Stubbe, 2012), translation tasks (Lemhöfer & Broersma, 2012; Zhang et al., 2019), and the Vocabulary Size Test (Nakata et al., 2020). In a comparison of four different task types for assessing vocabulary knowledge, (McLean et al., 2020) found that the Yes/No test had the highest internal reliability, took the least time to complete each item, and had correlations of 0.66 (SD 0.02) with reading proficiency, even at the shortest 5-minute test duration. However, it should be noted that these correlations at shorter test lengths were comparable, but slightly lower, than the other task types (0.69-0.76). Performance on the Yes/No test has also been shown to predict reading ability (Milton et al., 2010; Schmitt et al., 2011), as well as writing and listening abilities (Milton et al., 2010). Yes/No tests have also been used to assess vocabulary knowledge at various CEFR levels (Milton, 2010).

3.2 Fill in the Blanks

Fill in the Blanks shows a sentence with an unfinished word and asks test takers to complete the word, and by extension the sentence, using contextual cues (Figure 2). The unfinished portion of the word has a blank space for each character, indicating the length of the finished word and constraining possible responses. Test takers are administered six to nine *Fill in the Blanks* items (a sentence with an unfinished word) and have 20 seconds per item to respond.

⌚ 0:20 for this question

Complete the sentence with the correct word

I was just able to c a 1 a glimpse of the president as his motorcade passed by.

CONTINUE

Figure 2. A "Fill in the Blanks" item

Candidates for target words in *Fill in the Blanks* sentences are sourced from an English language corpus across four different parts of speech (adjectives, adverbs, nouns, and verbs) using lemma frequency. Sentences are then generated to provide the context that motivates the presence

of the candidate words across five different genres (literary, textbook, news, personal writing, and conversation), determined by the dispersion information from the corpus for the lemma. Additionally, we generate sentences with the target words in formulaic sequences, most notably collocations. Eligible collocates are determined using positional and syntactic co-occurrence data from the corpus. Sentences are also generated containing an antonym or synonym of the target word. The antonymy and synonymy are determined through a combination of WordNet (Princeton University, 2010) and the Moby Thesaurus List (Ward, 2002). Item parameters are estimated for each item using a feature-based 2PL model on response data. The 2PL model considers a number of features to estimate item difficulty parameters, including information about the sentence (e.g., the word frequency of the other words in the sentence and sentence-level semantic content features [McCarthy et al., 2021]).

Fill in the Blanks follows the format of the controlled-production vocabulary-levels test (PVLTL), a measure of controlled (as opposed to “free”) productive vocabulary knowledge that has been shown to distinguish high school students from university students in their performance across different word frequency levels (Laufer & Nation, 1999). *Fill in the Blanks* addresses both receptive and productive elements of vocabulary knowledge, including the form and meaning, concepts and referents, associations, grammatical functions, collocations, and constraints. The administration of *Fill in the Blanks* items ensures that test takers are administered items that tap into multiple aspects of vocabulary knowledge. It requires that all test takers receive at least one item containing an antonym or synonym of the target word as a context clue, and at least one item containing the target word in a collocation.

The construct coverage of *Fill in the Blanks* complements that of *Read and Select*, which targets primarily receptive knowledge. Collectively the two task types cover all aspects of the Nation model, as shown in Table 2. The two task types’ features also reflect the indicators of proficiency in vocabulary range and control from the CEFR. Specifically, the *Vocabulary range* scale includes the appropriate use of words in a context, appropriate use of collocations, and the ability to use synonyms. The *Vocabulary control* scale introduces the relationship between proficiency and lexical accuracy, as well as the increasing ability to use low-frequency vocabulary.

3.3 Vocabulary in Other Task Types

Vocabulary is one of the subconstructs in the DET’s writing and speaking tasks and, as such, is a key element in the **DET writing and speaking rubrics**. The *Duo Writing Scorer* and *Duo Speaking Scorer* each calculate numerous features that measure a learner’s vocabulary knowledge, from the proportion of words used per CEFR level to the degree of lexical accuracy and diversity in productive use, accounting for the breadth and depth of productive vocabulary knowledge.

The *Interactive Reading* task also includes vocabulary as a subconstruct with one of its five sub-tasks, namely the *Complete the Sentences* sub-task, wherein test takers read a short paragraph consisting of sentences with gapped words. For each gap, test takers select from a list of options the most appropriate word. This task type taps into the receptive level of the form and meaning aspect: identifying the meaning that each option signals and choosing the target word that best conveys the intended meaning.

Table 2. Aspects of vocabulary knowledge from Nation (2022) and how they correspond to Read and Select and Fill in the Blanks

Category	Aspect	DET task type	Corresponding task feature
Form	Written	<i>Read and Select</i>	Recognizing the written form when reading
		<i>Fill in the Blanks</i>	Using the correct spellings for the target word
	Word parts	<i>Read and Select</i>	Recognizing affixes and base words that are part of real English words
		<i>Fill in the Blanks</i>	Using the correct word parts for the target word in their appropriate forms
Meaning	Form and meaning	<i>Fill in the Blanks</i>	Recalling the target word to express the meaning inferred from the context
	Concepts and referents		Producing the target word to denote a particular meaning
	Associations		Knowing the synonyms and antonyms of target words (in some items)
Use	Grammatical functions	<i>Fill in the Blanks</i>	Recognizing the grammatical patterns in which the target word occurs and recalling the target word in its correct part of speech
	Collocations		Recognizing words that commonly occur with target words and recalling the target word as part of a collocation
	Constraints on use		Using target words in appropriate genres

In other task types, vocabulary is a secondary construct, where successful task completion depends less directly on a learner’s receptive and productive vocabulary knowledge. For instance, being able to recognize the word form when it is heard is crucial to successfully writing down an aural stimulus in the *Dictation* task type; similarly.

4 The e-Evidence Model

The e-Evidence model specifies evidence that links the observed performance to the latent construct (KSAs). There are two types of evidence; the first type is defining the rule for grading and scoring, or assigning numerical values to student performances, and the second type is showing how these values relate to the KSAs.

4.1 Grading and Scoring

Items on *Read and Select* and *Fill in the Blanks* tasks are assigned values according to binary grading (correct/incorrect). For each task type, the grades on several items are then combined into a score for the task type using 3PL (*Read and Select*) or 2PL (*Fill in the Blanks*) item parameters. Both task types are scored separately, with their item parameters also estimated separately. Items with poor model fit or low discrimination parameters are discarded. Thus, the acceptable model fit and item parameters of the operational items are one source of evidence that the items are tapping into a common latent construct.

4.2 Relationship to the KSAs

Over the course of one month, 3,000 *Read and Select* items and 600 *Fill in the Blanks* items were piloted on the practice test. On average, *Fill in the Blanks* items were harder than *Read and Select*, providing supporting evidence that *Fill in the Blanks* extends the range of inferences we can make about vocabulary knowledge from DET test tasks. Additional verbal protocol data corroborates the intended construct of *Fill in the Blanks* by showing that pilot participants were using their knowledge of synonyms, antonyms, collocations, and grammatical functions to respond to *Fill in the Blanks* items.

Disattenuated correlations estimated from repeater data showed that *Read and Select* and *Fill in the Blanks* correlated more highly with tasks assessing constructs that are theoretically relevant to vocabulary. More specifically, correlations with vocabulary-oriented task types (e.g., *Read and Complete*) were higher than those with less vocabulary-oriented task types (e.g., *Interactive Listening*), indicating that *Read and Select* and *Fill in the Blanks* are more likely to share KSAs (presumed to be vocabulary) with *Read and Complete* than they are with *Interactive Listening*[‡]. Additionally, reliability estimates for *Read and Complete* and *Fill in the Blanks* were within the acceptable range, comparable to those of other DET task types.

5 Conclusion

This white paper situates the development of vocabulary-focused task types on the DET within the extended evidence-centered design (e-ECD) framework, a component of the test's larger assessment ecosystem. The paper describes the DET's vocabulary knowledge construct aligned with established frameworks and models, how tasks are designed to elicit performance relevant to the construct, and psychometric modeling of task performance to target the construct. The integration of two task types that specifically measure vocabulary knowledge highlights the DET's continuous innovation as a language assessment. Specifically, it demonstrates how the DET leverages advances in AI technology to expand language proficiency construct coverage while increasing measurement efficiency.

[‡]We do not make explicit inferences about a test taker's vocabulary knowledge from their performance on the task types where vocabulary is a secondary construct.

References

- Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: research reviews* (pp. 77–117). International Reading Association. <https://files.eric.ed.gov/fulltext/ED203299.pdf>
- Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-eed) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in psychology, 10*, 441982. <https://doi.org/10.3389/fpsyg.2019.00853>
- Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: Some methodological issues in theory and practice. *Language Testing, 18*(3), 235–274. <https://doi.org/10.1177/026553220101800301>
- Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening, 14*(1), 14–31. <https://doi.org/10.1080/10904018.2000.10499033>
- Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test* (Duolingo Research Report No. DRR-22-01). Duolingo. <https://go.duolingo.com/ecosystem>
- Cardwell, R., Naismith, B., LaFlair, G., & Nydick, S. (2024). *Duolingo English Test: Technical manual* (tech. rep.). Duolingo. <https://go.duolingo.com/dettechnicalmanual>
- Clenton, J., & Booth, P. (2020). Introduction: Vocabulary and the four skills – current issues and future concerns. In J. Clenton & P. Booth (Eds.), *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 3–19). Routledge.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment – companion volume*. <https://www.coe.int/lang-cefr>
- Dabbagh, A., & Enayat, M. J. (2019). The role of vocabulary breadth and depth in predicting second language descriptive writing performance. *The Language Learning Journal, 47*(5), 575–590. <https://doi.org/10.1080/09571736.2017.1335765>
- Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors’ introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 1–32). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667268.003>
- Egbert, J., Burch, B., & Biber, D. (2020). Lexical dispersion and corpus design. *International Journal of Corpus Linguistics, 25*(1), 89–115. <https://doi.org/10.1075/ijcl.18010.egb>
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. <https://doi.org/10.1093/applin/amt015>
- González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41*(4), 481–505. <https://doi.org/10.1093/applin/amy057>

- Goodwin, S., Attali, Y., LaFlair, G. T., Runge, A., Park, Y., von Davier, A. A., & Yancey, K. P. (2023). *Duolingo English Test: Writing construct* (Duolingo Research Report No. DRR-22-03). Duolingo. <https://go.duolingo.com/scored-writing>
- Harrington, M., & Roche, T. (2014). Identifying academically at-risk students in an English-as-a-lingua-franca university setting. *Journal of English for Academic Purposes*, 15, 37–47. <https://doi.org/10.1016/j.jeap.2014.05.003>
- Hashimoto, B. J., & Egbert, J. (2019). More than frequency? exploring predictors of word difficulty for second language learners. *Language Learning*, 69(4), 839–872. <https://doi.org/10.1111/lang.12353>
- Langenfeld, T., Burstein, J., & von Davier, A. A. (2022). Digital-first learning and assessment systems for the 21st century. *Frontiers in Education*, 7. <https://doi.org/10.3389/educ.2022.857604>
- Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language testing*, 21(2), 202–226. <https://doi.org/10.1191/0265532204lt277oa>
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33–51. <https://doi.org/10.1177/026553229901600103>
- Lemhöfer, K., & Broersma, M. (2012). Introducing lextale: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, 44, 325–343. <https://doi.org/10.3758/s13428-011-0146-0>
- Masrai, A., & Milton, J. (2018). Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes*, 31, 44–57. <https://doi.org/10.1016/j.jeap.2017.12.006>
- McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-starting item parameters for adaptive language tests. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 883–899. <https://doi.org/10.18653/v1/2021.emnlp-main.67>
- McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting l2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, 37(3), 389–411. <https://doi.org/10.1177/0265532219898380>
- Meara, P. (1992). *Efl vocabulary tests*. Lognostics.
- Meara, P., & Buxton. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2), 142–154. <https://doi.org/10.1177/026553228700400202>
- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.
- Milton, J. (2010). The development of vocabulary breadth across the cefr levels. In I. Vedder, I. Bartning, & M. Martin (Eds.), *Communicative proficiency and linguistic development: Intersections between sla and language testing research* (pp. 211–232). Second Language Acquisition; Testing in Europe Monograph Series 1.
- Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *Eurosla monographs series* (pp. 57–78, Vol. 2). European Second Language Association.
- Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review*, 4(1), 151–172. <https://doi.org/10.1515/applirev-2013-0007>

- Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a Foreign Language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters. <https://doi.org/10.21832/9781847692900-007>
- Moghadam, S. H., Zainal, Z., & Ghaderpour, M. (2012). A review on the important role of vocabulary knowledge in reading comprehension performance. *Procedia – Social and Behavioral Sciences*, 66, 555–563. <https://doi.org/10.1016/j.sbspro.2012.11.300>
- Nakata, T., Tamura, Y., & Aubrey, S. (2020). Examining the validity of the LexTALE test for Japanese college students. *The Journal of AsiaTEFL*, 17, 335–348. <https://doi.org/10.18823/asiatefl.2020.17.2.2.335>
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
- Nation, I. S. P. (2013). *Learning vocabulary in another language (2nd ed.)* Cambridge University Press.
- Nation, I. S. P. (2022). *Learning vocabulary in another language (3rd ed.)* Cambridge University Press. <https://doi.org/10.1017/9781009093873>
- Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Duolingo English Test: Interactive reading* (Duolingo Research Report No. DRR-22-02). Duolingo. <https://duolingo-papers.s3.amazonaws.com/other/mpr-whitepaper.pdf>
- Park, Y., Cardwell, R., Goodwin, S., Naismith, B., LaFlair, G., Loh, K., & Yancey, K. (2023). *Assessing speaking on the Duolingo English Test* (Duolingo Research Report No. DRR-23-03). <https://duolingo-testcenter.s3.amazonaws.com/media/resources/speaking-whitepaper.pdf>
- Princeton University. (2010). About wordnet. <https://wordnet.princeton.edu/>
- Qian, D. D., & Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The routledge handbook of vocabulary studies* (pp. 66–80). Routledge. <https://doi.org/10.4324/9780429291586-5>
- Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, 7(2), 105–223. <https://revistas.um.es/ijes/article/view/49021>
- Read, J., & Dang, T. N. Y. (2022). Measuring depth of academic vocabulary knowledge [Advance online publication]. *Language Teaching Research*. <https://doi.org/10.1177/13621688221105913>
- Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, 64(4), 913–951. <https://doi.org/10.1111/lang.12077>
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95, 26–43. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Settles, B., T. LaFlair, G., & Hagiwara, M. (2020). Machine learning-driven language assessment. *Transactions of the Association for computational Linguistics*, 8, 247–263. https://doi.org/10.1162/tacl_a_00310
- Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, 29(4), 471–488. <https://doi.org/10.1177/0265532211433033>

Ward, G. (2002). Moby thesaurus list. <https://www.gutenberg.org/ebooks/3202>

Zhang, X., Liu, J., & Ai, H. (2019). Pseudowords and guessing in the Yes/No format vocabulary test. *Language Testing*, 37(1), 6–30. <https://doi.org/10.1177/0265532219862265>