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

February 2017 – Volume 20, Number 4

An Academic Definitions Test

Daniel Richard Isbell
Michigan State University, USA
<isbellda@msu.edu>

Abstract

L2 vocabulary is commonly conceptualized in terms of a size or depth of one's total lexical knowledge and tested discretely with selection-type items. Concerns exist, however, regarding L2 users' ability to cope with unknown vocabulary, especially in the context of academic reading. This motivated the creation of a test which measures the ability to identify definitions for unknown terms in an academic text. The test comprised a 768-word introductory biology textbook excerpt with non-words replacing ten technical terms, and test takers were asked to write definitions for each term. The test was piloted with 158 prospective university students of varying L2 English proficiency. Support for test score interpretations was sought by investigating the test's reliability, correlation with reading comprehension, and the qualities of test-taker responses. Internal consistency was adequately high, and a moderate correlation with reading comprehension was found. Test-taker responses aligned with expectations, demonstrating utilization of the in-text definitions. The test shows promise for those interested in academic reading abilities and vocabulary learning.

Keywords: academic reading, English for academic purposes, lexical familiarization, vocabulary learning, vocabulary testing

Introduction

In academic contexts, second-language (L2) users are required to read texts for a variety of purposes, including comprehension, synthesis, and evaluation. Another purpose of academic reading, if not as overt, is to become familiar with key vocabulary in a particular discipline (Flowerdew & Peacock, 2001), a task that can be especially challenging in an L2 (Gablasova, 2014). While popular academic reading tests focus on comprehension and, to perhaps a lesser extent, synthesis and evaluation (e.g., Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt & Schedl, 2000; Moore, Morton & Price, 2012), less attention is paid specifically to discipline-specific vocabulary (e.g., lower representation of low-frequency vocabulary in the IELTS reading module compared to university textbooks [Green, Ünaldi & Weir, 2010]). When vocabulary items are embedded in popular academic reading tests (e.g., TOEFL), they are generally few in number and often answerable based on pre-

existing knowledge. Vocabulary-focused testing has been mainly concerned with the existing size and depth of a learner's vocabulary (Read, 2007; Schmitt, 2008), with very few tests measuring abilities to form knowledge of new words in context.

University textbooks, a primary source of academic reading material, include considerable amounts of low-frequency technical vocabulary (Chung & Nation, 2003). It has been suggested that these terms are beyond the realm of L2 vocabulary instruction (Read, 2000). This creates an obvious problem for L2 readers, who often find themselves falling short of knowing the 98% of words in a text necessary for full comprehension (Schmitt, Jiang & Grabe, 2011). However, textbooks also have pedagogic aims, and often provide readers with lexical familiarizations through in-text definitions (Bramki & Williams, 1984; Selinker, Trimble, & Trimble, 1976) or at times a glossary of technical terms at the end of the book. If we accept the suggestion that technical vocabulary is beyond most L2 vocabulary instruction, utilizing in-text definitions effectively would seem to be an important skill for both successful text comprehension and advanced vocabulary growth and retention.

Recognizing the limits of conventional vocabulary instruction and measures of vocabulary size, scholars have made calls to move towards considering vocabulary in specific contexts for teaching (Hyland & Tse, 2007) and testing (Read, 2007; Read & Chapelle, 2001). In response, the current article describes the development and initial steps toward validation of the Academic Definitions Test (ADT), which measures an ability to identify definitions of technical vocabulary in the specific context of a university textbook. Immediately following are brief reviews of research on learning of vocabulary during reading and vocabulary testing, and then an overview of the ADT design is presented.

Background

Learning Vocabulary through Reading

As mentioned, vocabulary knowledge is considered extremely important to reading comprehension, and in turn researchers have also examined the role of reading in the acquisition of vocabulary. In this area, research on lexical inferencing, or how readers are able to infer the meaning of unknown words in a text, is most relevant. Nassaji (2006) found that learners are generally not very successful at determining the meaning of new words in a text, though relations among lexical inferencing success, depth of lexical knowledge, and quality of strategy use were found. Zahar, Cobb, and Spada (2001) found repetitions of new, unknown vocabulary to be important to learning, noting that the repetitions provided a broad range of contextual supports. Both Nassaji and Zahar et al.'s research used what could be considered highly general reading passages – a news article and a fable written for children, respectively – and asked learners to determine the meaning of 10 to 30 unknown (or likely to be unknown) words. Other studies have focused on tertiary learners in academic contexts. Hamada (2009), in case studies of Japanese ESL students reading English for academic purposes pedagogic texts, found that instruction in lexical inferencing strategies was effective, and that the students moved from local to global strategies over time. Kaivanpanah and Alavi (2008) found that Iranian university learners of English differed in their ability to infer meanings according to their

overall proficiency and, importantly, the complexity of the text. These findings support Nation's (2013) suggestion that context may provide useful syntactic and discourse cues that help learners associate words with their meanings. While this research into learning vocabulary through reading is of obvious interest to the ADT, very little work appears to focus on the meanings of words for which in-text definitions are given in an academic text. Rather, research tends to emphasize contextual clues and strategies for successful utilization of the clues, usually in a general or somewhat controlled text. A notable and recent exception is Gablasova (2014), which investigated the learning of technical terms in the L1 and L2 via academic texts with lexical familiarization, and found that L2 readers were less successful in learning the terms and retaining knowledge of the terms in a delayed post-test.

Vocabulary Testing

In his 2007 overview of vocabulary assessment, John Read noted that measuring learner vocabulary size has been the primary focus of most vocabulary tests. Tests such as the Productive Vocabulary-Size Test (Laufer & Nation, 1999), Vocabulary Size Test (Nation & Beglar, 2007), Vocabulary Levels Test (Nation, 2001), and the Y_Lex test (Meara & Miralpeix, 2006) all share the common characteristic of discrete, low-context items that aim to measure test taker knowledge of the meanings of words, often using corpus-derived frequency information to extrapolate total vocabulary size. By comparing scores on a vocabulary test with scores on skills-oriented proficiency tests, research on vocabulary size tests has also shown positive correlations between vocabulary size and language proficiency (Laufer & Nation, 1999; Harrington & Carey, 2009). While measuring vocabulary size in terms of meaning knowledge has been popular in vocabulary testing for some time, some efforts to explore depth of vocabulary knowledge have also emerged. These efforts include Read's (1998) Word Associates Test, measuring semantic and collocational knowledge of words, and Paribakht and Wesche's (1997) Vocabulary Knowledge Scale, which attempted to measure depth of knowledge through a self-assessment/discrete item hybrid format.

Vocabulary items included in reading skills tests, such as the TOEFL iBT reading section or the Pearson Test of English Academic, also tend to rely on existing vocabulary knowledge, even though the word (or blank) is presented in context. The following TOEFL example item illustrates this point (ETS, 2015, p. 4):

> The word "excavating" on line 25 is closest in meaning to
> a. digging out
> b. extending
> c. destroying
> d. covering up

Readers who are unfamiliar with the word excavate may benefit from contextual clues (e.g., a crater is mentioned in the sentence), but those who are familiar with the word would not need to utilize context at all. Thus, scores on vocabulary size tests as well as many academic reading tests would not facilitate interpretations about the ability to understand unknown words in context.

Lastly, while not as often discussed in vocabulary testing literature, a popular means of testing productive vocabulary is embedded in constructed-response academic speaking and writing tasks (e.g., the TOEFL speaking and writing subtests) in which lexical variety and accuracy/appropriateness are rubric criteria used to determine an overall speaking or writing proficiency. Generally, lexical sophistication (e.g., diversity, richness, use of formulaic sequences) in written or spoken test tasks is correlated with higher scores (e.g., Crossley, Salsbury & McNamara, 2012; Read & Nation, 2006; Yu, 2009). This consideration of vocabulary certainly speaks to the role of vocabulary in the academic domain, but does not address the understanding of unfamiliar vocabulary.

Taken together, vocabulary testing research has shown that an L2 user's vocabulary (receptive knowledge and productive diversity) correlate positively with general language abilities, but less is known about the interaction of unknown vocabulary and skills in specific contexts. One vocabulary test that does involve understanding new words in a written text is Sasao's (2013) Guessing from Context Test (GCT). The GCT input employs non-words in short narrative texts, and has three sections of multiple choice questions targeting parts of speech, contextual clues, and meaning. The latter of the three, which requires test takers to derive the meaning of a non-word in a short text, shares some test design features with the ADT (described in detail next). However, like the instruments used in research on lexical inferencing, the GCT focuses primarily on contextual clues rather than intentional lexical familiarization (e.g., in-text definitions).

Design of the Academic Definitions Test

The ADT seeks to make an inference about an aspect of vocabulary and reading ability given little attention: a test-taker's ability to recognize definitions in an academic text. A high-ability examinee should be able to read through academic texts and recognize textually-provided definitions for key terms, which is a common, if not overt, purpose in many academic reading contexts (Bramki & Williams, 1984; Flowerdew & Peacock, 2001; Gablasova, 2014; Selinker, Trimble, & Trimble, 1976).

The ADT comprises one reading passage with 10 dichotomously scored short-answer items (Appendix A). The reading passage was excerpted from a textbook used in a university's 100-level biology course (*Life: The Science of Biology* (7th ed.), Purves, Sadava, Orians & Heller, 2004). The excerpt was minimally edited for length (768 words) in order to have ten key terms supported by in-text definitions. The decision to base the test on an extended, coherent passage was made in order to better reflect the domain of academic reading. While only sampling the biology domain, it is worth noting that introductory biology courses are commonly required of many undergraduate science majors and can be used to satisfy general studies requirements for other majors in many American universities. Additionally, the definitions present in the text generally align with definition structures reported by Bramki and Williams (1984, based on an economics textbook) and Selinker, Trimble, and Trimble (1976, based on university-level scientific and technical texts). For example, Selinker, Trimble and Trimble (1976, p. 284) cite the following definition of *barometer* (source unknown):

> *A barometer is a meteorological instrument used for the measurement of atmospheric pressure.*

Compare with the following definition for metabolic rate in the biology textbook (Purves et al., 2004, p. 962):

> *The metabolic rate of an animal (see Chapter 41) is a measure of the overall energy needs that must be met by the animal's ingestion and digestion of food.*

Both examples could be said to follow a typical academic definition structure, including the term (barometer, metabolic rate), semantic class (meteorological instrument, a measure) and distinguishing details (what each is used for). This structure is also described by Xiang and Grabe (2007) as a prevalent discourse structure, one of about 12-15 that appear across a wide variety of texts. Other terms in the ADT are elaborated on by information provided over the course of a paragraph, comparable to Selinker, Trimble & Trimble's (1976) account of paragraphs with *description* or *classification* rhetorical purposes. Bramki and Williams (1984) also note explicit definitions being commonly used to familiarize readers with new lexical items, and add *exemplification* and *explanation* as techniques for familiarization, both of which are found in the ADT. Furthermore, they noted that virtually all instances of in-text lexical familiarization in an economics textbook were targeted at nouns; this is reflected in the ADT lexical items, all of which are single nouns or noun phrases (adjective+noun or noun+noun).

In the ADT, the target technical terms were replaced with non-words. In the case of multi-word technical terms, the second (and generally less technical) word was unchanged (e.g., *budget* in *energy budget* was unchanged). An effort was made to avoid introducing potentially confusing morphological information. For example, the word *physical* in *physical activity* was replaced with *kerepal*, retaining the *–al* suffix which allows the word to be identified as an adjective. The decision to obscure the core semantic meanings of the original technical vocabulary was made in order to control for existing vocabulary knowledge, thereby increasing the context-dependence (Read & Chapelle, 2001) of the items. In other words, items were constructed in such a way that would prevent results from being confounded by test-takers simply happening to know a term or making guesses that did not consider the in-text support. Item-by-item explanations for how each technical term was altered is presented in Table 1.

Test-takers responded to items by producing short definitions ranging from one word to a sentence or two, and direct quotes were allowed. In the target domain, this type of response might be found in student notes or even formal written assignments. Responses were considered correct if they captured the "core meaning" (gist) of the word in the context of the passage; insufficient responses are considered incorrect. A sample test item and relevant input is presented in Figure 1 to illustrate (see also Appendix A). The sample, used in the ADT directions, is from a different text and the original term is *sustainable behaviors*.

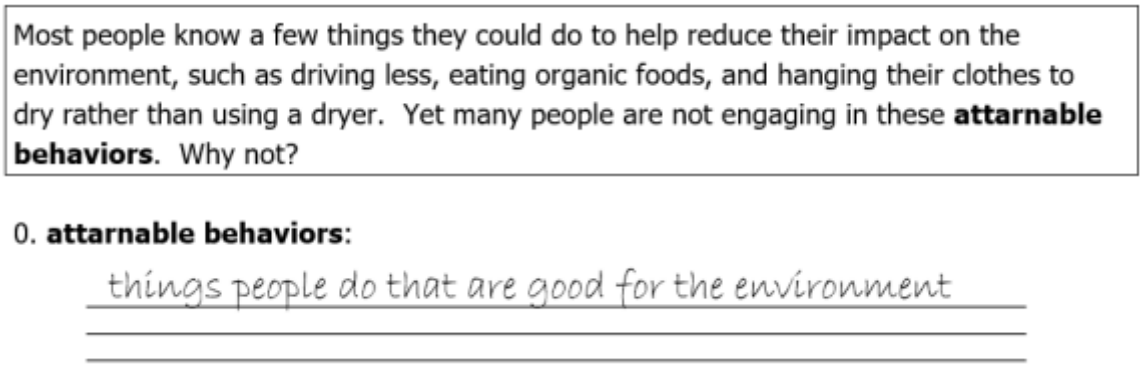

***Figure 1*. Example ADT item**

Table 1. ***Technical Vocabulary and Non-word Substitutions***

Original Vocabulary Item (part of speech)	**Non-word Substitution**	**Explanation**
1. heterotroph (n)	telpon	Both morphemes substituted.
2. autotroph (n)	calpon	This item required a connection to item 1. Item 1's final morpheme *pon* was duplicated.
3. calorie (n)	moffen	Entire word substituted; *calorie* is common but the text provides a precise technical definition.
4. metabolic rate (adj. + n phrase)	geradic rate	First word substituted, *-ic* morpheme retained to maintain adjective status. Second word unchanged, as the general sense of *rate* applies.
5. physical activity (adj. + n phrase)	kerepal activity	First word substituted, *-al* morpheme retained to maintain adjective status. Second word unchanged, as the general sense of *activity* applies.
6. energy budget (n + n phrase)	sceltel budget	First word substituted as the definition relates to a specific, technical sense of *energy*. Second word kept, as the general sense of *budget* still applies.
7. fat (n)	miresa	Entire word substituted. *Fat* has a precise technical meaning in the text.
8. undernourishment (n)	subtrinement	The first two morphemes were substituted, while the third was retained to maintain noun status. The first morpheme

Original Vocabulary Item (part of speech)	Non-word Substitution	Explanation
		was changed to a semantically-similar morpheme.
9. kwashiorkor (n)	viottis	Entire word substituted.
10. overnourishment (n)	extratrinement	The first two morphemes were substituted, while the third was retained to maintain noun status. A connection to item 8 was necessary, and *extra*, loosely similar to *over*, was chosen.

Purpose of the Study

The goal of this study is to provide preliminary support for the argument that ADT scores are interpretable as an indication of the ability to identify definitions of unknown technical vocabulary in academic texts. Evidence for the scoring, generalization, and extrapolation inferences necessary for test score interpretation is considered (Kane, 2013).

To support the scoring inference in an interpretive argument, primary backing often comes from a well-designed scoring key and set of procedures (Kane, 2013). Additionally, test-taker responses can be investigated for match-up with expected responses. If there are fundamental mismatches between test-taker responses and expected responses, the suitability of the key and scoring procedure would be seriously jeopardized. Further backing for the claim that scores are reflective of performances can be provided by item statistics. An item with poor statistics may have "technical defects" (Miller, Linn, & Gronlund, 2009, p. 353) in its construction or scoring method. For example, a short answer item with a vague key would be difficult to score consistently and may have a poor item discrimination statistic as a result. On the other hand, an item with desirable statistics could provide some supplementary evidence of correspondence between actual and expected responses.

The second inference, generalizability, is generally supported by information on the reliability of the test (Kane, 2013). High internal consistency of scores is desired for both tests and research instruments (Hatch & Lazaraton, 1991), and allows for more confident generalization of test interpretations beyond a given test form and sample. Another consideration for generalizability when subjectivity is present in scoring is inter-rater agreement (Landis & Koch, 1977). High inter-rater agreement would support the notion that test interpretations would be generalizable with little regard for who scores a given set of responses.

Finally, the extrapolation inference involves the relationship between test scores and performance in real-world contexts. Kane (2013, p. 28) describes two primary types of evidence for the extrapolation inference: analytic and empirical. Analytic evidence is provided in the test design and its relation to the target domain (see previous section).

Also, if test-taker processes are reasonably aligned with the processes of the task in general, it can be taken as evidence supporting the extrapolation of test scores. Empirical evidence can involve the investigation of overlap with a criterion measure that samples from the same domain.

To investigate the validity of ADT interpretations, the following research questions (RQs) were addressed:

1. Do items perform as intended?
2. Do test-taker responses demonstrate that the ADT measures what is intended?
3. Is the ADT a reliable measure?
4. What is the relationship between the ability to recognize definitions in texts and general reading comprehension ability?

RQ1 and 2 relate to the scoring inference. For RQ1, it was expected that test items would have suitable difficulty and discrimination values based on previous pilots of the ADT. For RQ2, it was hoped that test takers would indeed respond to the task by producing, or clearly attempting to produce, definitions of the non-word terms. RQ3 deals with the generalizability inference. Though the length of the test is short (K=10), previous piloting suggested that reliability would be acceptably high. For RQ4, addressing the extrapolation inference, there was expected to be a moderate degree of correlation between reading comprehension and the ability to identify definitions in texts based on shared domain sampling. However, due to the unique design of the ADT, a large correlation was also not expected.

Methods

Participants

Participants were a group of 158 test takers who took a complete placement battery at an American university's intensive English program (IEP) in August 2013. These test takers represented both sexes (male = 95, female = 63) and included mainly L1 Arabic (n = 37), Chinese (n = 80), and Portuguese (n = 24) speakers, with smaller numbers of Japanese, Korean, French, Spanish, and Bengali speakers. Their overall English proficiency ranged from low (placed into the beginner levels of the IEP) to high (admitted as full-time university students). These test-takers represented a larger population of adult L2 English speakers from other countries who have the goal of studying in the US. The sample's test forms were archived by the IEP and made available to the researcher for the purposes of this study.

Instruments

Academic Definitions Test. The present form of the ADT was piloted during the IEP's Summer 2013 placement battery and was greatly refined from an initial pilot conducted in March of 2013. The 10-item Summer 2013 pilot was taken by 13 test-takers, with a mean of 3.23, a standard deviation of 2.65, reliability of 0.80, and a standard error of measurement (SEM) of 1.18.

Responses were scored by the researcher using a key. Additionally, a second rater was trained by the researcher for the purpose of investigating reliability. This training included a review of the task and scoring guide (Appendix B) and a norming session covering five sets of responses.

Reading subtest. The IEP's reading subtest was used to make inferences about a test takers' reading comprehension abilities. Reading comprehension was defined as the ability to understand written English, which included understanding main ideas and details, making reasonable inferences, and identifying the relationship among ideas in a text. The IEP's Fall 2013 reading subtest featured 40 dichotomously scored multiple-choice (4 option) items grouped into five testlets ordered by increasing difficulty. The internal consistency (Cronbach's alpha) of the reading subtest was .81.

Procedures

Analyses. Two variables were examined in this study. First and foremost, the ability to recognize definitions in texts was considered. The ability to recognize definitions in texts was operationalized as a score on the ADT, with scores ranging from 0 to 10. Reading comprehension was another important variable considered in this study, operationalized by a score on the IEP placement battery's reading subtest, with scores ranging from 0 to 40. Summary statistics and histograms were examined for these variables.

To investigate item performance (RQ1), item difficulty (P) and item discrimination (point-biserial, D) values were calculated. RQ2, involving the alignment of test taker responses with expected responses, was analyzed qualitatively. Test taker responses for each item were first sorted by correctness. Then, incorrect responses were again sorted based on content similarities, allowing themes to emerge. A descriptive label was created for each incorrect response theme (e.g., *incomplete*, *only examples*). For each item, 2-3 themes were sought; this was thought to provide adequate description without losing interpretability in the context of a 10-item test.

To answer the question of test reliability (RQ3), scores were analyzed by examining the item-level responses of all test-takers and computing Cronbach's alpha (RQ3.1). Additionally, due to a degree of subjectivity present in scoring the ADT's short-answer responses, inter-rater agreement was also examined (RQ3.2). The agreement of scores assigned by two raters for 25 test-takers (15.8% of the total) was computed via Cohen's kappa. Cohen's kappa is a measure of agreement for dichotomous variables that accounts for chance, making it more conservative than most agreement estimates (Landis & Koch, 1977).

RQ3, concerning the relationship between ADT and reading comprehension scores, was answered by testing the null hypothesis that there was no relationship between the two sets of scores. The relationship was examined by Pearson product-moment correlation.

Administration. Groups of test takers were assigned to different rooms for the placement battery, which was proctored by IEP faculty. All test takers took the reading subtest (60 minutes) and the ADT (20 minutes) on the same day, and the ADT always followed the reading subtest. Each group had a different schedule for the battery, with slightly staggered start times, where subtests were taken in different order to facilitate lab use for the speaking subtest.

Results

This section includes descriptive statistics for the ADT and the Reading Subtest, and the results of analyses addressing each RQ.

Descriptive Statistics

Descriptive statistics of the Reading Subtest and ADT scores are found in Table 2.

Table 2. ***Descriptive Statistics of the Reading Subtest***

	N	K	Min	Max	Mean	SD	Reliability	SEM
ADT	158	10	0	10	5.12	2.84	.79	1.33
Reading Subtest	158	40	1	34	18.88	6.51	.81	2.84

Note: Reliability = Cronbach's alpha.

For the ADT, the complete range of scores was used and the mean (5.12) was near the median (5). With a mode of 8 and a standard deviation of 2.84, the resulting distribution is somewhat platykurtic and negatively skewed (see Figure 2). Nonetheless, normality is assumed for subsequent analyses and for the interpretation of test scores; subsequent analyses (item statistics and correlations) are either fairly robust to violations of normality or do not require it.

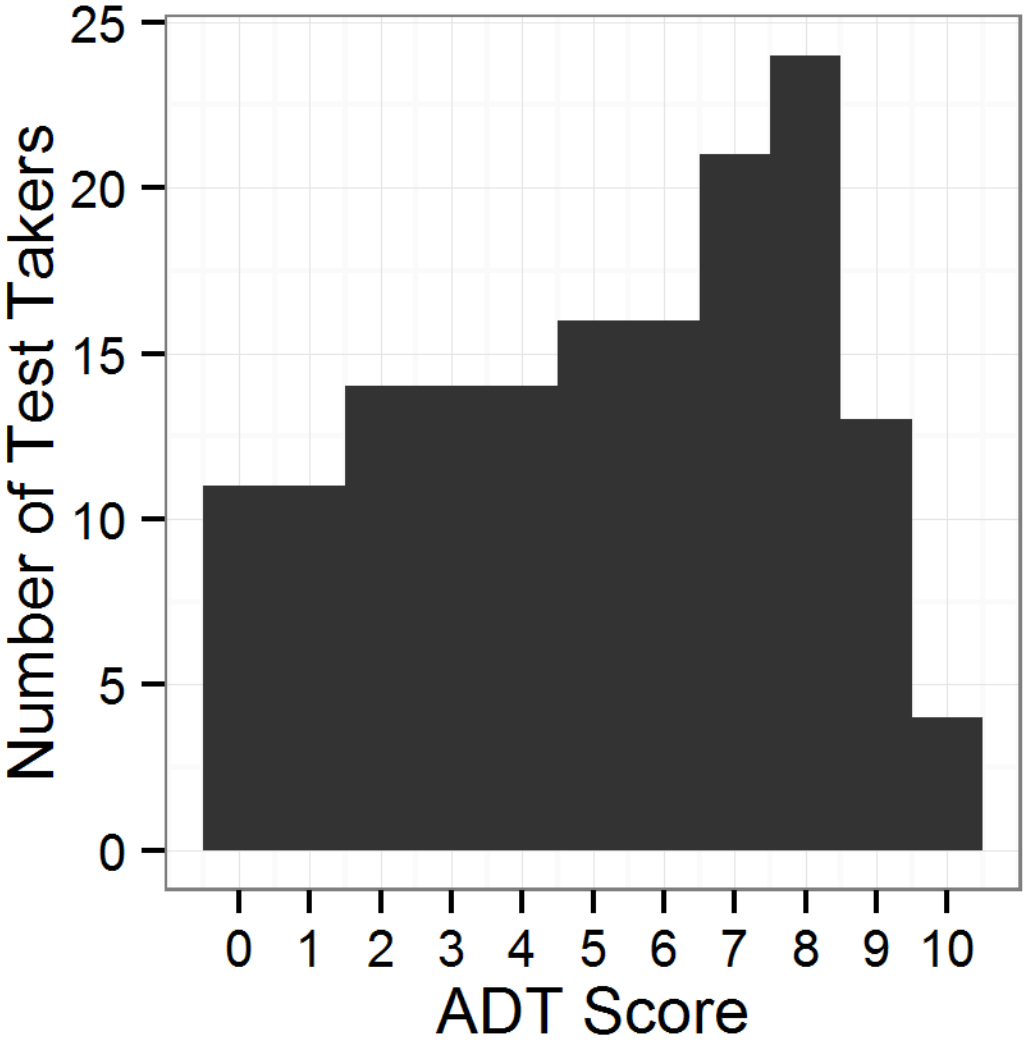

***Figure 2*. Distribution of ADT scores**

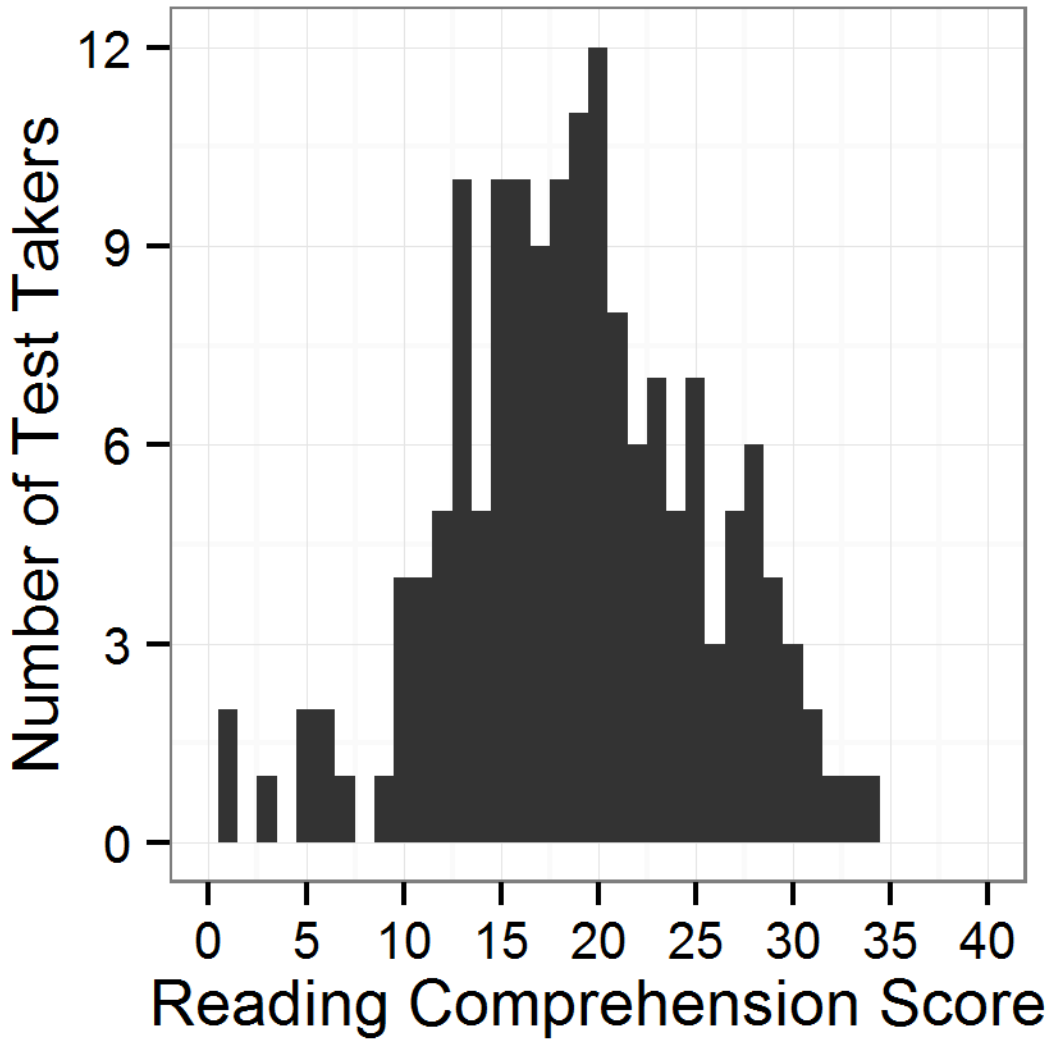

***Figure 3*. Distribution of reading comprehension scores**

Research Questions

RQ1. Item statistics as well were computed for the ADT to address RQ1, which asked whether items performed as intended. Item statistics, including difficulty (P) and discrimination (point-biserial, D) are provided in Table 3, with means at the bottom of the table. Difficulty ranged from .34 (greater difficulty) to .65 (less difficulty), with a mean of .51 for the whole test. Discrimination values ranged from .28 to .63 with a mean of .46.

Table 3. ***ADT Item Statistics***

Item	P	D
1	0.65	0.45
2	0.53	0.45
3	0.64	0.46
4	0.61	0.51
5	0.35	0.28
6	0.34	0.40
7	0.69	0.63
8	0.47	0.46
9	0.37	0.53
10	0.46	0.45
Mean	**0.51**	**0.46**

Note. P = item difficulty; D = item discrimination.

RQ2. RQ2 asked whether test taker responses matched expectations for an in-text academic definitions task. Table 4 describes and exemplifies themes in test taker responses (exemplars are presented in quotes and italics). For each test item, the key and an exemplar correct answer is provided. For incorrect responses, 2-3 themes were identified and exemplified in the rightmost column. As can be seen, successful responses could be extremely close to the key, demonstrating a strong match-up with expected and actual responses. Among incorrect responses, many themes included part of the definition, or perhaps were too vague or imprecise, yet still somewhat on the right track. In Item 1 (*telpon*, or heterotroph), for example, one incorrect response theme included the idea of eating, but contained some imprecisions, such as "it's animals eat meat" (heterotrophs include more than just carnivores). Other incorrect response themes involved a confusion due to unrelated or misinterpreted information in the immediate context. For Item 5 (*kerepal* activity, or physical activity), some test takers provided basal energy requirements as a definition, not quite picking up that physical activity requires energy beyond homeostasis. Many test takers borrowed words, phrases, or whole clauses/sentences from the text. Some test takers left items blank. In sum, most incorrect responses demonstrated solid attempts to define the non-words, and shortcomings often demonstrated partial understanding of the term.

Table 4. ***Test Taker Correct and Incorrect Responses***

Item	Key and Exemplar	Incorrect Response Themes and Exemplars
1. telpon	animals that get energy from eating other organisms *"the animals that derive nutrition from other organisms"*	No mention of eating organisms *"Animal that's depend on something for their food."* Includes eating, but vague or imprecise *"It's animals eat meat."* Wrong/confused defintion *"things animals use as energy sources"*
2. calpon	organisms that get energy from the sun *"the calpon can production your energy by photosynthesis"*	Only examples *"single cell organisms"* Herbivore *"the name of the animal just eat plant"* Innaccurate *"the animals or plants eated by other animals"*
3. moffen	a measure of heat, heat necessary to raise 1g water 1℃ *"a unit we can use to measure heat energy"*	Incomplete or insufficient *"a units"* Digestion *"the animal eat the food and turn it into part of itself"* Conflated with animals or energy budget *"energy necessary for the activity"*
4. geradic rate	an animal's total energy need *"the energy an animal need to get per day"*	Vague or imprecise *"the rate about the energy the animals ingested or digested"* Incomplete *"it's a measure of the energy"*
5. kerepal activity	work done with one's body *"some activity like for person doing labor…"*	Sedentary *"a special thing for people doing sedentary work"* Basal energy requirements *"this basal energy requirement"* Inaccurate or vague *"kind of energy, something that can use energy"*
6. sceltel budget	a comparison of calories (moffens) consumed with calories spent *"is difference between moffens consumed and moffens expended"*	Incomplete, focusing on analysis OR energy *"a kind of budget that can help people to analyze the cost which is cost-benefit"* Diet or health *"a plan to improve the health"* Food, or the caloric content of food *"number the moffenic value of any food an animal eats"*
7. miresa	a type of stored energy in the body *"the important form of*	Incorrect process *"it's part of body that make energy for all body"* Incorrect class *"things can be stored*

Item	Key and Exemplar	Incorrect Response Themes and Exemplars
	stored energy in the bodies of animals"	*water"* Vague or incomplete *"good for energy"*
8. subtrinement	when an animal takes in less energy than required *"take too little food, and it can't meet its energy requirements"*	Only about energy requirements *"energy requirements for the animal"* Result of subtrinement *"take the reserve of necessary energy in body"* Eating less (neutral or positive) *"animals need little food to meet its energy"*
9. viottis	a syndrome caused by undernourishment (subtrinement) in which proteins are broken down for energy *"name of a syndrome, using own proteins for fuel"*	Stages before or unrelated to kwashiorkor *"there syndrome that protein loss is minimized for as long as possible"* Incomplete or imprecise *"a syndrome that is resulted by subtrinement"* Incomplete, only results *"the animal will be death"*
10. extratrinement	when an animal takes in more energy than required *"take more food than its requirement and store as body fat"*	Example or purpose only *"the bear use this to save energy to use when hibernating"* Vague, missing the "what" *"demand is less than supply"* Fat (or another effect) *"a serious health hazard"*

RQ3. RQ3 was broken down into two subquestions. RQ3.1 asked whether scores on the ADT were internally consistent. Cronbach's alpha was used to investigate this question, and it assumes data are nominal or ordinal and contain variance. Table 3 provides evidence of item variance through item difficulty and discrimination, satisfying the assumptions of Cronbach's alpha. The overall internal consistency for the ADT was .79. While this value is quite respectable for a 10-item test, the Spearman-Brown predictive reliability formula was used in a post-hoc analysis to estimate reliability if 5 more items were added, resulting in a value of .85. RQ3.2 asked whether scores on the ADT were consistent between raters. Cohen's kappa was used to investigate this question, and it assumes that classifications are subjective and accounts for chance agreement (typically yielding lower coefficients than other agreement analyses). Given that ADT items are human scored, and in this case were scored by two human raters, the assumptions of the procedure were considered satisfied. Average inter-rater agreement for the ADT was .79 and kappa values for individual items ranged from .54 to 1 (agreements for each item are found in Table 5).

Table 5. ***ADT Inter-rater Reliability***

Item	1	2	3	4	5	6	7	8	9	10	Mean
kappa	1	.54	.60	.68	.72	.65	.92	.82	1	1	.79

Note. Kappa = Cohen's kappa.

RQ4. RQ4 asked if there was a relationship between ADT scores and Reading Subtest scores. A Pearson product-moment correlation was used to determine the relationship between the two scores. Pearson correlations assume equivalent reliability (ADT = .79, Reading Subtest = .81), independent data on continuous scales, normal distributions (see Descriptive Statistics above), and a linear relationship (see Figure 4); all assumptions were considered to have been met. Pearson product-moment correlation between the two measures yielded a statistically significant result of .52 ($N = 158$, $r_{critical} = 0.19$, $p < .001$). The strength of association (R2) for this relationship was .28. Due to the suspect normality of the ADT score distribution, Spearman's rho was also computed, yielding a value of .53 ($p < .0001$).

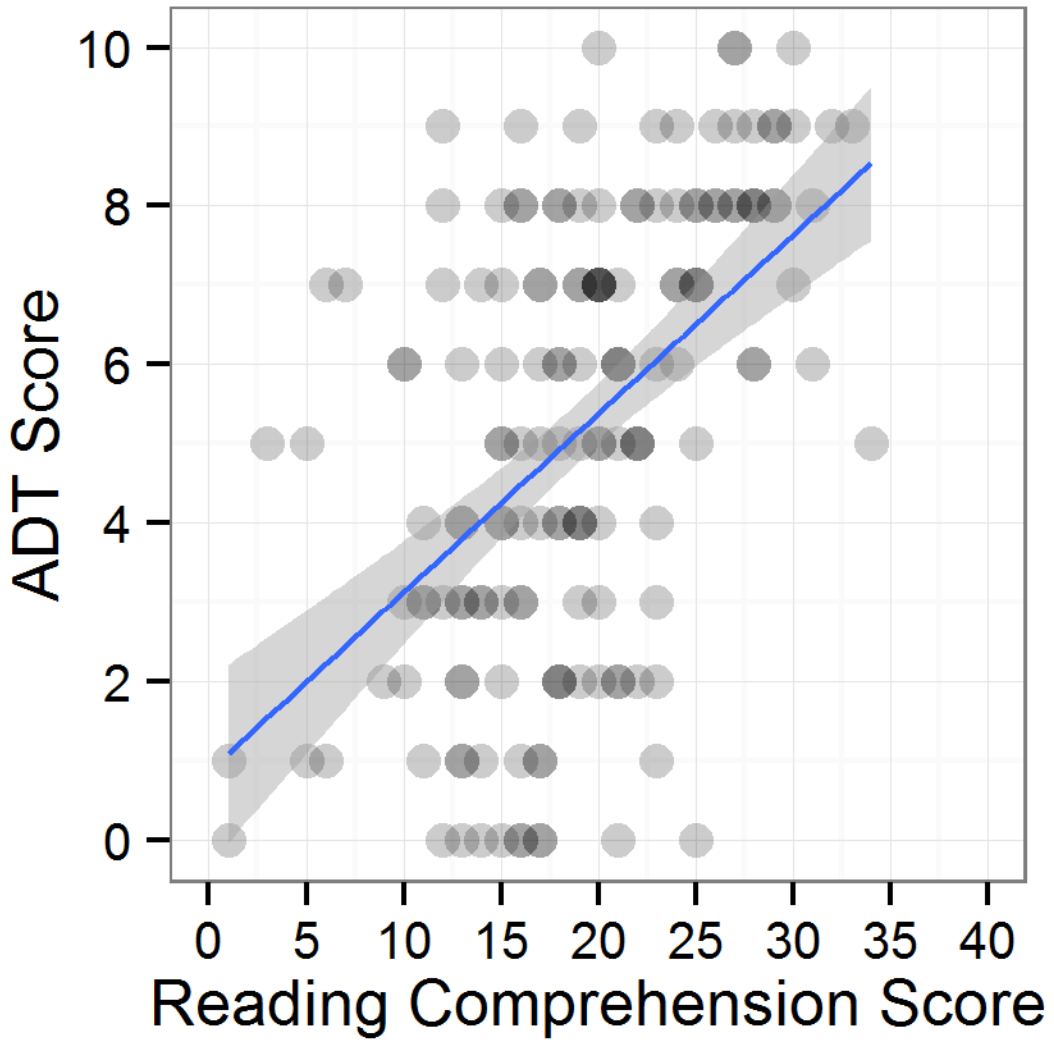

***Figure 4*. Relationship between reading comprehension and ADT scores**

Discussion

To summarize, descriptive statistics of the ADT were found to adequately support norm-referenced test interpretations, meaning that ADT items performed well, with item statistics falling within desirable ranges for norm-referenced tests (RQ1). Responses demonstrated understanding of the task and clear signs of interaction with the text (RQ2). The ADT was found to be adequately reliable (RQ3). Furthermore, there was a relationship between reading comprehension and the ability to identify definitions in a text (RQ4), though the two measures were not highly overlapping. These findings, along with the design of the test and scoring procedures, provide initial evidence for the scoring, generalization, and extrapolation inferences necessary for meaningful interpretation of ability based on ADT scores.

Analysis of ADT items and responses provided valuable evidence for the scoring inference. Individual items had difficulties that fell within a suitable range (i.e., .25-.75) for norm-referenced interpretations and all items had desirable discrimination values. The items, then, appear to work as intended within the scoring scheme, with no items being extremely hard or easy, or otherwise surprising (e.g., low-scorers doing better on a particular item than high-scorers).

Analyzing test taker responses also provided evidence for the scoring inference. Primarily, many test-takers provided answers that were very close to keys. Incorrect answers could often be described as misinterpretations, too vague, or otherwise missing key elements of the definitions. The results also suggested that guessing was rather limited, and that test takers were utilizing the context provided by the textbook excerpt, as there was very little meaning they could glean from the non-word items alone. These findings also support the extrapolation inference, in that the interaction with the in-text definitions appears reasonably similar to how one would identify technical vocabulary definitions (and perhaps take notes) in the target domain across texts and disciplines. Many test takers chose to incorporate quotations for part or all of their definitions, demonstrating meaningful interaction with the text. Even unsuccessful responses, such as those featuring incomplete definitions, often incorporated words or ideas from the text. In sum, the task appeared to be highly transparent to test takers, and appeared to elicit definitions based on in-text information.

The generalizability inference was supported by internal consistency and inter-rater agreement. Internal consistency was high, though not quite surpassing .8, which would have been desirable. A low number of items is likely a limiting factor. In comparison, the much longer Vocabulary Size Test (K = 90, Laufer & Nation, 1999) reported reliabilities of .86 and .91 for two different forms. Adding five more items to the ADT would yield an estimated reliability of .85. Inter-rater agreement (.79) can be interpreted as substantial (Landis & Koch, 1977). With that substantial level of inter-rater agreement, scores from one rater appear sufficient and suggest that the test and scoring guide engender limited subjectivity to scoring decisions. Some individual items, however, had lower inter-rater agreement. The two items with the lowest agreement, items 2 and 3, do not seem to share any particular characteristic that would cause lower agreement. In the future, efforts should be made to revise the scoring guide and rater training procedures for any use of the ADT.

Evidence supporting the extrapolation of test scores was also found. ADT scores had a statistically significant, moderately positive correlation with reading subtest scores (r = .52), which was intuitively expected, as the ADT is thought to use syntax and discourse knowledge associated with general reading comprehension to understand where definitions begin and end. However, the correlation was not high enough to say that the tests measured highly overlapping abilities (R^2 = .28); the remaining variation suggests that the ADT measures something different from general reading comprehension. This difference may involve strategy use, the quality of which has been found to influence lexical inferencing success (Hamada, 2009; Nassaji, 2006), or awareness of definition-related discourse structures. In sum, the ADT had some degree of overlap with general

reading comprehension, but also measures a unique ability to recognize definitions in a text.

Conclusion

Limitations of this study primarily include threats to internal validity. While the sample was generally representative of those who come to the US to pursue university studies, it was ultimately a convenience sample. The instruments were found to have acceptable and nearly equal reliability, but higher reliabilities would have led to greater confidence in results. These factors impose some limits on the generalizability of ADT score interpretations.

Additionally, the design of the ADT has potential drawbacks. The inclusion of only 10 items was unfortunate, but ultimately a practical necessity: test-takers attempted the ADT during the administration of a long (~4 hour) placement battery, and granting any more time for the ADT was determined to be detrimental. Naturally, a longer test would have yielded more robust information regarding test takers' abilities. Other design decisions may have weaknesses, too. For example, the ADT input only represents one text from one academic discipline. This allowed for a reasonably high degree of task authenticity, with multiple appearances of each term in a rich, coherent text for test takers to utilize. However, this approach limits the representation of the academic domain, and may have advantaged some test takers due to topic familiarity (Clapham, 1996). Perhaps a more analytic test format, with shorter excerpts across several disciplines, would be desirable. Another alternative would be to integrate the ADT task into a traditional academic reading test, with ADT-like items embedded in the testlets. These alternatives are worth pursuing in future work exploring in-text definitions.

This study has implications for both researchers and language testers. The ADT measures a little-studied ability to identify definitions for unknown technical terms in a text and the present results provide some support for score interpretation, providing a potentially valuable tool for researchers interested in academic reading, reading strategies, discourse knowledge, lexical inferencing, and notions of vocabulary beyond an underlying trait of existing word knowledge. For testers, results suggest that the ADT draws on abilities and/or knowledge not represented in general reading comprehension tests, highlighting an element of reading for academic purposes that may be worth investigating and embedding into high-stakes academic reading tests, such as those used for university admissions. Similarly, information from the ADT could be useful for placement into an academic language program and informing reading instruction. Any of these potential uses, however, would require the collection of additional supporting evidence.

While the ADT and the results of this study have provided new insights into an interesting intersection of academic reading and vocabulary, the components of this ability and how it might be taught to learners require further investigation. Comparing ADT results to a traditional receptive vocabulary test may answer the question of how much of this intersection draws on existing vocabulary knowledge. Corpus studies may be able to shed light on the most common patterns used in academic texts to provide definitions, which would have immediate implications for instruction and for revisions of the ADT. Finally,

more thoroughly investigating how learners successfully identify definitions in a text could lead to a better understanding of this ability and provide further pedagogical implications.

About the Author

Daniel R. Isbell is a PhD student in Second Language Studies at Michigan State University. His main research interest is language assessment, with a focus on assessments for academic or specific purposes. He is also interested in L2 vocabulary teaching and learning.

References

Bramki, D., & Williams, R. C. (1984). Lexical familiarization in economics text, and its pedagogic implications in reading comprehension. *Reading in a Foreign Language, 2*, 169-181. Retrieved from http://nflrc.hawaii.edu/rfl

Carkin, S. (2001). *Pedagogic discourse in introductory classes: Multi-dimensional analysis of textbooks and lectures in biology and macroeconomics* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Publication number: AAI3004014).

Chung, M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language, 15*, 103-116. Retrieved from http://nflrc.hawaii.edu/rfl

Clapham, C. (1996). The development of IELTS: A study on the effect of background knowledge on reading comprehension. *Studies in Language Testing, 4*. Cambridge: Cambridge University Press.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243-263. doi: 10.1177/0265532211419331

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P, Mulcahy-Ernt, P., & Schedl, M. (2000). TOEFL 2000 Reading Framework: A Working Paper. *TOEFL Monograph Series MS 17*. Princeton, NJ: Educational Testing Service.

ETS (2015). *TOEFL iBT test questions*. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf

Flowerdew, J., & Peacock, M. (2001). The EAP curriculum: Issues, methods, and challenges. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp.167-194). New York, NY: Cambridge University Press.

Gablasova, D. (2014). Learning and retaining specialized vocabulary from textbook reading: Comparison of learning outcomes through L1 and L2. *Modern Language Journal, 98*(4), 976-991. doi:10.1111/modl.12150

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd edition). Harlow, UK: Longman.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing, 27*(2), 191-211. doi:10.1177/0265532209349471

Hamada, M. (2009). Development of L2 word-meaning inference while reading. *System, 37*, 447-460. doi:10.1016/j.system.2009.03.003

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System, 37*, 614-626. doi:10.1016/j.system.2009.09.006

Hatch, E., & Lazaraton, A. (1991). *Design and statistics for applied linguistics: The research manual*. Boston, MA: Heinle & Heinle Publishers.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*, 235-253.

Jiang, X., & Grabe, W. (2007). Graphic organizers in reading instruction: Research findings and issues. *Reading in a Foreign Language, 19*, 34-55. Retrieved from http://nflrc.hawaii.edu/rfl/

Kaivanpanah, S., & Alavi, S. M. (2008). The role of linguistic knowledge in word-meaning inferencing. *System, 36*, 172-195. doi:10.1016/j.system.2007.10.006

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50* (1), 1-73. doi:10.1111/jedm.12000

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174. doi:10.2307/2529310

Laufer, B., & Nation, I. S. P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*, 33-51. doi:10.1177/026553229901600103

Meara, P. M., & Miralpeix, I. (2006). *Y_Lex: The Swansea advanced vocabulary levels test*, v2.05. Swansea: Lognostics.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

Moore, T., Morton, J., & Price, S. (2012). Construct validity in the IELTS Academic Reading test: A comparison of reading requirements in IELTS test items and in university study. *IELTS Research Reports, 11*, 1-89.

Nassaji, H. (2006). The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success. *The Modern Language Journal, 90*, 387-401.

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

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

Paribakht, T.S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huck (Eds.), *Second language vocabulary acquisition* (pp. 174-200). Cambridge, UK: Cambridge University Press.

Purves, W., Sadava, D., Orians, G., & Heller, H. (2004). *Life: The science of biology* (7th ed.). Sunderland, MA: W. H. Freeman and Company.

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

Read, J. (2000). *Assessing vocabulary*. New York, NY: Cambridge University Press.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7*, 105-125. doi:10.1017/S0261444812000377

Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*, 1-32. doi:10.1177/026553220101800101

Read, J., & Nation, I.S.P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports, 6*, 207-231.

Sasao, Y. (2013). Frequently asked questions (GCT). Retrieved from: http://ysasaojp.info/VocabTests/GCT/GCT_FAQ.pdf

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329-363.

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. doi:10.1111/j.1540-4781.2011.01146.x

Selinker, L., Trimble, T. R. M., & Trimble, L. (1976). Presuppositional rhetorical information in EST discourse. *TESOL Quarterly, 10*, 281-290.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*, 236-259. doi:10.1093/applin/amp024

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *The Canadian Modern Language Review. 57*, 541-572. doi:10.3138/cmlr.57.4.541

Appendix A

Academic Definitions Text

Vocabulary in Context

Directions: Read the passage. Then, write a definition for each of the 10 **bolded terms** (a term may be a single word or a short phrase). You do not need to write complete sentences. You may use your own words. You may use words from the passage.

Example:

Most people know a few things they could do to help reduce their impact on the environment, such as driving less, eating organic foods, and hanging their clothes to dry rather than using a dryer. Yet many people are not engaging in these attarnable behaviors. Why not?

0. attarnable behaviors:

things people do that are good for the environment

Appendix B
Scoring Guide

Scoring: 1 for acceptable (captures the core meaning of the term), 0 for unacceptable (does not express core meaning) based on the following definitions. Acceptable answers may be as short as one word, or may be longer. Direct quotation is allowed.

Nutrient Requirements

Animals must eat other organisms to stay alive. Since they derive their nutrition from other organisms, they are called **telpons**. In contrast, **calpons** (most plants, some bacteria, and some protists) trap solar energy through photosynthesis and use that energy to synthesize all of their components. Directly or indirectly, **telpons** take advantage of – indeed, depend on – the organic synthesis carried out by **calpons. Telpons** have evolved an enormous diversity of adaptations to exploit, directly or indirectly, the resources made available through the actions of **calpons**. In this section, we will discover how animals use those resources as energy sources and as building blocks for complex molecules.

Energy can be measured in moffens

In chapter 6, we learned that energy in the chemical bonds of food molecules is converted to provide animals with energy for cellular work. This conversion is inefficient, however; in fact, most of the energy that was in the food is lost as heat. Therefore, we can talk about the energy requirements of animals and the energy content of food in terms of a measure of heat energy: the **moffen**. One of these units is the amount of heat necessary to raise the temperature of 1 gram of water 1◦C. Since this value is a tiny amount of energy compared with the energy requirements of many animals, physiologists commonly use the kilomoffen (kmof) as a unit of measure (1kmof = 1,000 moffens).

The **geradic rate** of an animal (see Chapter 41) is a measure of the overall energy needs that

(1-10) Write a definition for each term below:

1. telpon:

_____

_____

_____

2. calpon:

_____

_____

_____

3. moffen:

_____

_____

_____

4. geradic rate:

_____

_____

_____

5. kerepal activity:

must be met by the animal's ingestion and digestion of food. The basal **geradic rate** of a human is about 1,300-1,500 kmof/day for an adult female and 1,600-1,800kmof/day for an adult male.**Kerepal activity** adds to this basal energy requirement. For a person doing sedentary work, about 30 percent of total energy consumption is due to **kerepal activity**, and for a person doing heavy physical labor, 80 percent or more of total moffenic expenditure is due to **kerepal activity**.

It is possible, of course, to quantify the moffenic value of any food an animal eats. It is also possible to quantify the moffenic cost of anything an animal does. By comparing moffens consumed with moffens expended, it is possible to construct **sceltel budgets** for any set of circumstances. Sceltel budgets allow ecologists and evolutionists to apply a cost-benefit analysis to any behavior.

Although the cells of the body use energy continuously, most animals do not eat continuously. Therefore, animals must store fuel molecules that can be released as needed between meals. Carbohydrates are stored in the liver and muscle cells as glycogen, but the total glycogen store represents only about a day's basal energy requirements (1,500-2,000 kmof). **Miresa** is the most important form of stored energy in the bodies of animals. Not only does **miresa** have more energy per gram than glycogen, but it can be stored with little associated water, making it more compact. Migrating birds store energy as **miresa** to fuel their long flights; if they had to store the same amount of energy as glycogen, they would be too heavy to fly! Proteins are not used as energy storage compounds, although body protein can be metabolized as an energy source of last resort.

Subtrinement and Extratrinement

If an animal takes in too little food to meet its energy requirements, it is **subtrined**, and must make up the shortfall by metabolizing some of the molecules of its own body. This consumption

_____

_____

_____

6. sceltel budget:

_____

_____

_____

7. miresa:

_____

_____

_____

8. subtrinement:

_____

_____

_____

9. viottis:

_____

_____

_____

10. extratrinement:

_____

_____

of self begins with the energy storage glycogen and **miresa**. Protein loss is minimized for as long as possible, but eventually a starving animal begins to break down its own proteins for fuel. The syndrome that results is called **viottis**. Blood proteins are among the first to be used, resulting in loss of fluid to the intercellular spaces (edema; see Chapter 49). Additional consequences of protein deficiency are breakdown of the immune system and degeneration of the liver. Muscles waste away, and eventually even brain protein is lost, leading to mental retardation. If starvation continues, the breakdown of body proteins eventually leads to death.

When an animal consistently takes in more food than it needs to meet its energy requirements, it is **extratrined**. The excess nutrients are stored as increased body mass. First, glycogen reserves build up; then additional dietary carbohydrates, fats, and proteins are converted to body fat. In some species, such as hibernators, seasonal **extratrinement** is an important adaptation for surviving periods when food is not available. In humans, however, **extratrinement** can be a serious health hazard, increasing the risk of high blood pressure, heart attack, diabetes, and other disorders.

Text adapted from: Purves, W., Sadava, D., Orians, G., & Heller, H. (2004). Life: The science of biology (7 th ed.). Sunderland, MA: W. H. Freeman and Company.

_____

THIS IS THE END OF THE READING TEST.

Appendix B

Scoring Guide

Scoring: 1 for acceptable (captures the core meaning of the term), 0 for unacceptable (does not express core meaning) based on the following definitions. Acceptable answers may be as short as one word, or may be longer. Direct quotation is allowed.

1.	**telpon** (heterotroph)	animals that get nutrition/energy from eating other organisms/living things
2.	**calpon** (autotroph)	organisms that get energy from the sun/ organisms that use solar energy for nutrition
3.	**moffen** (calorie)	a unit of measure that tells the amount of heat necessary to raise the temperature of 1 gram of water 1 degree Celsius / a measure of heat
4.	**geradic rate** (metabolic rate)	an animal's total energy need
5.	**kerepal activity** (physical activity)	work people do with their bodies
6.	**sceltel budget** (energy budget)	for any animal, a comparison of moffens (calories) consumed with calories expended
7.	**miresa** (fat)	stored energy in the body
8.	**subtrinement** (undernourishment)	when an animal takes in too little food to meet its energy requirements
9.	**viottis** (kwashiorkor)	a syndrome caused by subtrinement (undernourishment) in which proteins are broken down for fuel

10.	**extratrinement** (overnourishment)	when an animal takes in more food than it needs to meet its energy requirements