The CATESOL Journal

BRENNA SHEPHERD
University of California, Irvine

Vocabulary Assessment With Varying Levels of Context: A Replication Study

This replication study investigates how the level of context in vocabulary assessment affects the scores on tests of American idioms. Using Uçkun's methodology of 3 tests with 3 levels of context, 85 participants varying in level from high-beginner to advanced took an online test consisting of 30 questions, 10 questions for each level of context. The tests were matching, sentence-level gap filling, and rational-delete cloze. The participants were nonnative speakers of English living in the US. The scores were analyzed for mean differences and in regard to differences in native language and English proficiency level. No significant differences were seen for native language, but there were significant differences for context level and proficiency level. As an exploratory part of the study, 17 of the participants were asked to perform a think-aloud protocol task while taking the test. Their responses were recorded and analyzed descriptively for insight into test-taking strategies.

Introduction

With the rise of international tests of English proficiency and the increasing numbers of English as a foreign language (EFL) speakers and English as a second language (ESL) speakers, a need has arisen to create tests that accurately assess the skills of students. In the world of foreign language learning, assessment is often at the forefront of discussion because assessment is the only measure that teachers, administrators, and students have to gauge a student's proficiency. With the recent trends toward alternative assessments and macro-skills versus discrete-skills testing, the question arises as to what subcomponents of language are necessary to assess. Reading comprehension, writing ability, and speaking and listening ability, as part of the core learning curriculum, are accepted as necessary testing fields. But vocabulary, as a micro skill, has received less attention in both language teaching and assessment. Vocabulary acquisition has been linked to success in reading (McQueen, 1996;

Qian, 2008), writing (Arnaud, 1992; Laufer & Nation, 1995), and general language proficiency (Meara & Jones, 1988). As Wilkins (1972) says, "While without grammar very little can be conveyed, without vocabulary *nothing* can be conveyed" (p. 111). Thus, vocabulary is the elemental form of communication and should be treated as a necessary part of language learning, and as such, at some point, vocabulary will need to be assessed.

Since vocabulary seems to be so elementary to so many language skills, the question then arises as to how educators should be testing vocabulary as a means of economically indirectly testing other skills. Vocabulary can be tested in many ways, with varying amounts of context. In recent years, there has been a push to test vocabulary in context because concerns about positive washback, or backwash (i.e., "the effect that tests have on learning and teaching" [Hughes, 1989, p. 53]), are on the rise (Read, 2007). The communicative language teaching method is replacing more grammar-based, decontextualized teaching and as a result, researchers are calling for tests to match teaching methods. Yet research on the matter is still incomplete, and only a handful of studies have empirically tested how much context is appropriate (Qian, 2008; Uçkun, 2008).

Uçkun's study (2008) found that in vocabulary assessments that used different levels of context, statistically significant differences in test scores were seen between tests with no context (matching) and texts with high context (rational-delete cloze) for some groups, but there was no significant difference for matching and sentence-level gap filling. Her study, conducted in an EFL context, is valuable to the field of vocabulary assessment but still leaves some questions unanswered. If researchers in the field of vocabulary assessment are pushing for a communicative approach to testing based on the fact that this type of testing is more related to teaching methods and therefore more accessible to students, then there should be significant differences across all groups in regard to increased context, with an increase in score as context increases. Because this is not the case with Uçkun's study, more research is required. The present study attempts to replicate Uçkun's original study with modifications that explore a more in-depth understanding of the most common strategies students use when faced with a vocabulary test.

This study seeks to add further research to the field of assessing vocabulary in context, specifically in relation to how much context is needed, how that context is used by the test taker, and whether context is more or less useful for speakers of different skill levels and languages. This study tested three levels of context (matching, sentence-level gap filling, and rational-deletion cloze). The participant population

was ESL students studying in the US, with a minimum proficiency level of high-beginner. The study is focused most specifically on testing situations such as the Test of English as a Foreign Language (TOEFL), in which ability to inference, in addition to present vocabulary level, is integral to the testing process. Through a series of tests and a think-aloud protocol task, information was gathered to answer questions of how participants process vocabulary in tasks with different levels of context.

Literature Review

What Does It Mean to "Know" a Word?

As Mezynski (1983) said: "Word meanings can be 'known' to varying degrees. Depending on the task, a person could perform adequately with relatively imprecise knowledge. In other situations, a much finer notion of the word's meaning might be required" (p. 265). Research also suggests that the number of words a person recognizes is far greater than the number of words a person can actually use. Nation (2001) has said that knowing a word involves "subknowledges," which include the morphological (form: spoken, written), the syntactical (collocations, constraints on use including register and frequency), and the semantic (meaning, including form and meaning, concept and reference, and associations). Clearly, with so many nuances of a word, testing what it means to know a word can be difficult.

Many researchers talk about breadth and depth when discussing lexical knowledge (Qian, 1998, 1999; Read, 1989; Wesche & Paribakht, 1996). Breadth is, in simple terms, the size of the vocabulary (i.e., the mental lexicon). This means that knowledge of a word could be superficial, but that a person is able to recognize the word. Depth, on the other hand, is how well a subject knows a single word. Laufer (2004) tested four types of knowing: active recall, passive recall, active recognition, and passive recognition. Active recall involves providing a word when a definition is given. Passive recall is providing a definition when a word is given. Active recognition is choosing a target word from a list of words when a definition is given, and passive recognition is choosing the correct definition of a word when a list of definitions is given. In her test, Laufer found that the passive mode is easier than the active mode and that recognition is easier than recall. She also found that higher-frequency words were easier to identify than lower-frequency words, as defined by Wu, Adams, and Wilson's ConQuest software (1998).

Despite the complex layers of vocabulary, most ESL teachers are still stuck on the one-word, one-meaning way of testing. Students rarely have to manipulate word families, morphemes, or identify mul-

tiple meanings or connotations for words (Folse, 2004). Even after the grammar-translation method came under attack by the communicative method, the focus continues to be shifted away from a lexical approach (other than Michael Lewis's work, 1993, 1997). Vocabulary teaching has always received less focus than grammar, and so the research on the subject is still lacking.

Types of Language Testing

Despite the lack of explicit vocabulary teaching, explicit vocabulary testing has been popular for many years because of the ease with which one can be tested in vocabulary versus the other macro skills. Assessing a 30-question multiple-choice test in vocabulary is much simpler than assessing a 400-word essay. Assessment in language learning exists for several reasons. While alternative forms of testing have come into recent prominence, the standardized test (i.e., mostly multiple choice) is still seen as the most common and possibly most affordable type of testing. With so much focus on testing, the questions still remain as to the best format for testing.

Traditionally, vocabulary was tested using a discrete, direct-translation or multiple-choice format, in which vocabulary items were listed and students were required to translate the items into or from a native language or to choose from a possible list of synonyms or definitions (Read, 2007). Because there are multiple meanings and constructions of many words in the English language (Taylor, 1998), students may not know which of these meanings to choose from. The multiple-choice test format provides a limited sampling of a learner's knowledge, and learners may choose the right answer by process of elimination, which again is an inaccurate estimation of knowledge.

Notwithstanding the popularity of multiple-choice vocabulary testing, alternative methods exist. Now, as the communicative approach to teaching gains popularity, new types of testing are being developed that mirror teaching methods. Such methods include the cloze test and the C-test, gap filling, sentence-writing items, word associates testing, and matching (Read, 2000).

Henning (1991) looked at the TOEFL in regard to eight multiple-choice format tests to determine whether familiarity with testing type would affect performance on the tests and whether all eight reliably tested the same thing. The eight types were:

1. Word given in a sentence with subsequent multiple-choice (MC) synonyms;
2. Isolated word/phrase with MC synonyms;

3. Minimal sentence stem matched to MC synonyms;
4. Minimal sentence stem cloze with MC options for cloze blank;
5. Reduced-length inference-generating stem (i.e., more context than minimal) with MC options;
6. Reduced-length interference-generating stem cloze;
7. Single word/phrase embedded within a sentence with MC options for each embedded item; and
8. Single word/phrase embedded within an extended reading passage.

He found that familiarity did not affect performance and that the tests did reliably test the same thing. However, he did note that the only alternative method investigated that outperformed the current (as of 1999) testing method (method 1 above) in reliability was the test that embedded items in a reading passage.

As the TOEFL is the most recognized test of English language proficiency for students who want to study at the university level in the US, this test may be one of the most important types of tests to examine. According to the TOEFL website, the TOEFL "measures your ability to use and understand English at the university level. And it evaluates how well you combine your listening, reading, speaking and writing skills to perform academic tasks" (ETS, 2012). Elemental to this idea is that the TOEFL is not just interested in how you perform right now, but also how you will perform in the future throughout your university career in the US. Vocabulary tests, which have been linked to success in the four major skills areas listed above, are a part of the TOEFL test. Because the TOEFL is interested in not just your immediate vocabulary level, it seems essential that the TOEFL also test your ability to inference and use context clues to arrive at knowledge of previously unseen vocabulary.

Tying the TOEFL to alternative types of assessments, Qian (2002) did a study in which he measured the importance of both vocabulary knowledge depth and vocabulary size in relation to performance on basic reading comprehension for the TOEFL 2000. As one of the first studies of its kind, Qian's research was limited to partial dimensions of vocabulary depth (synonymy, polysemy, and collocation). Qian gave three tests—Reading for Basic Comprehension-TOEFL, Depth-of-Vocabulary-Knowledge Measure (DVK) (Read, 1993), and the Vocabulary Levels Test (Nation, 1983)—to 217 students enrolled in the Intensive English Program at Toronto University. He found that the DVK and vocabulary size measure different aspects of vocabulary

knowledge, but that they are equally important to predicting reading-comprehension abilities. Thus, it is possible that the methods that the TOEFL uses may need to add some of the more alternative types of testing.

Summary of the Original Study

As mentioned before, very little research has been done regarding assessment of vocabulary items in context. In her 2008 study, Uçkun's main goals were to find out if changing the amount of context surrounding the assessed vocabulary words would create significantly different results and to decide if different proficiency levels responded differently from each other. Her research design included testing three complete classes (189 participants)—intermediate, upper-intermediate, and advanced levels—of EFL speakers from a Turkish university. The tests she used varied in amount of context given, from an isolated matching, a semicontextualized sentence-level gap-filling test, to a contextualized rational-deletion cloze test. The same vocabulary words were tested on all three tests, but tests were changed depending on the students' level. Two tests for each type of test were created, with 10 discrete words tested on each test. A total of 20 words were therefore tested at each level. To choose the words on the tests, Uçkun consulted the teachers of the classes and also Nation's Range and Frequency programs (Heatley, Nation, & Coxhead, 2002) to analyze the number of words on her tests, which were in the most frequent 1,000, 2,000, and 3,000 words in the English language. Her purpose for this comparison was to determine whether her vocabulary for the passages for the cloze tests was too low or too high for the group's level. As a result of the variance of words tested, Uçkun did not make comparisons among the groups, but rather within the groups. Using one-way analysis of variance (ANOVA), Uçkun found that for the advanced and intermediate groups, the means on the three tests were significantly different, but the means on the upper-intermediate group's tests did not significantly differ, although gap-filling tests did seem to receive the highest score with the highest reliability. For intermediate and advanced levels, the difference in the cloze from the matching and gap filling differed significantly, but there was no significant difference between matching and gap-filling tests. Table 1 shows the means for each group and test.

Thus, contrary to prior assumptions, this study corroborates Qian (2008) by suggesting that there is no significant difference between matching and fill-in-the-blank tasks. And, contrary to the view that contextualized tests help students perform better, all levels performed

Table 1
Uçkun (2008) Results

	Matching	*Gap filling*	*Cloze*
Advanced	11.92	10.47	5.86
Upper-intermediate	10.15	10.41	9.69
Intermediate	12.23	12.59	6.59
Average total	11.43	11.16	7.38

better on gap-filling assessments rather than the more contextualized cloze passages.

Moreover, the highest overall score was for matching questions, then gap filling, and then cloze. These findings are very contradictory to the idea that context *helps* students. Cloze tests saw test scores nearly three points lower than for matching and gap filling.

Because research of this nature requires further study to be generalized, especially in different populations, and because Uçkun's study did not answer the question of *how* context is used, the present research study will attempt to replicate Uçkun's findings for a different population. As Uçkun did not answer to what extent students use context to help them decide on the correct answer, this study will address that gap by asking questions in relation to how students manipulated their answers based on the given contextual clues. This study is also more interested in tests on a greater scheme, meaning general tests of vocabulary that could be made of all levels, rather than tests that are designed for a specific classroom. The results of this study are meant to provide more information on how to create proficiency tests.

Contextualized Tests and the Ability to Inference

Because it has been established that the TOEFL is interested in reasoning skills as a secondary motive in testing, some time needs to be spent on what researchers have found in regard to how second language (L2) learners process context as compared to L1 learners. Studies on native speakers have shown that young readers are able to use the context clues to help them figure out the meanings of unknown words. Nagy, Herman, and Anderson's study (1985) on eighth graders showed that those who had read a passage before completing a vocabulary test performed significantly better than those who took the test with no reading passage.

In the case of L2 learners, studies have shown that L2 learners may not be able to pick up on meaning based on context. Deighton (1959) pointed out that context does not reveal the meanings of words

as often as is assumed. In fact, as Marks, Doctorow, and Wittrock (1974) point out, "Unfamiliarity with low frequency words, perhaps with only one such word in a sentence, may render meaningless an entire sentence, which may, in turn, inhibit comprehension of the meaning of subsequent sentences in the same passage" (p. 262). In the case of nonnative speakers, it has been shown that contextualized clues are not as readily used or recognized. Laufer (1987) argues that learners must know as much as 95% of the vocabulary in a passage to even begin to use the contextual clues, and Schatz and Baldwin (1986) found that ESL students did not perform well in tasks requiring them to guess the meanings of words from context. Laufer and Ravenhorst-Kalovski (2010) further corroborated these findings in a study in which they analyzed the connection between reading comprehension and vocabulary. They found that there was a slight increase in reading comprehension as vocabulary increased, but that there are two thresholds.

Although it has been shown that L2 learners may have more trouble using context clues to discern the meaning of new vocabulary, Nation (1990) says that learners can be taught strategies for learning low-frequency words rather than being taught the words explicitly. Because low-frequency words are so numerous (several hundred thousand as compared to two to three thousand high-frequency words) and because they occur so infrequently, teaching low-frequency words can be unnecessary. Once learners have mastered the top three thousand most frequent words, they should be able to infer meanings, but ESL students have to be taught how to infer meanings, unlike native speakers. For example, Nation tested the level of inference on an untaught class. Achievement ranged from 0-80%. After the class was taught how to infer, the achievement range increased to 50-85%.

If teaching strategies to students can mitigate their inability to deduce meaning from context, then using context is a skill. Skills can be learned, as is the case here. As a skill, using context can be tested. I would argue that testing this skill on the TOEFL is necessary as it is something that will be used over and over again at the university level. The ability to infer is an important skill, and by testing vocabulary in context, the test performs a dual-task: measuring vocabulary knowledge and the ability to infer.

Idioms

Because idioms were used as the test items for the present study, some time will be spent in explaining the nature of idioms and how idioms were chosen as the test item of choice. Swinney and Cutler (1979) identify an idiom as "in its simplest form … a string of two

or more words for which meaning is not derived from the meanings of the individual words comprising that string" (p. 523). Goldberg (2003) says that idioms are a type of construction, unpredictable based on its component parts, which functions like any other conventionally recognized lexical item. Street et al. (2010) classified idioms from the American National Corpus (ANC) into three types: verb-noun constructions, prepositional phrases, and subordinate clauses. They began with 4,500 sentences from the ANC: a third from written nonfiction, a third from transcribed written narratives, and a third from written fiction. Annotators tagged the idioms according to the three types, and the completed list was used for the present study.

Because idioms are often not acquired in a classroom setting, idioms can be used to test another dimension of language learning—implicit. Unless participants have taken a class that specifically teaches idioms, participants are not likely to have received extensive instruction in idioms. Idioms receive little focus in the language classroom as compared to other, more formal skills. Because the present study does not intend to test vocabulary knowledge, but rather how students use context to figure out vocabulary terms during assessment, it is less relevant that participants know the vocabulary items being tested. Therefore, the present study may be used as evidence for contextualized tests' also testing the ability to make inferences.

Research Questions

1. Does increasing the amount of context from no context (matching) to reduced context (sentence-level gap filling) to extended context (rational-deletion cloze test) show significantly different results on computerized tests of American idioms?
2. Does the participant's English language proficiency level and native language influence his or her ability to answer questions correctly?
3. How do participants use the context given to help them decide the meanings of unknown vocabulary terms?

Methodology

Participants

A total of 85 participants were included in the study, and 17 of them were part of a think-aloud protocol task. Their responses were included in the overall statistical tests as well as analyzed separately for additional information regarding use of deduction strategies. Most participants were Arabic native speakers (75%). Proficiency levels of

the participants were mostly based on the class level in which they were enrolled at California State University, Long Beach's American Language Institute (ALI). For those participants who did not attend the ALI, level was based on the number of years a participant had been in the US and the number of years he or she had studied English. While the researcher recognizes that this is not a perfect system, no other option was available given the reduced likelihood that participants would take both a vocabulary test and a proficiency test without major incentives. The proficiency level of only 10 participants had to be determined in this way. The researcher is also aware that those 10 participants work daily in an English-speaking environment.

Materials

After a pilot study involving 23 ESL participants and 9 native speakers was conducted, an online test using Google Docs was created. The test had a total of six sections: Biographical Information, Matching Questions, Sentence Questions, Cloze Questions, Opinion Questions, and Prizes. The Biographical Information section asked questions about the participants' gender, native language, and time spent studying English. Because Nation (2001) suggests at least 30 items for vocabulary tests, the test had 30 items broken into three sections with 10 questions each. These sections were modeled after Uçkun's (2008) study (Matching, Sentence, Cloze), for which matching questions had no context, sentence gap-filling questions were one to two sentences, and the cloze was a complete paragraph with introduction, body, and conclusion. In each section, a drop-down menu was created so that participants could choose from 11 different answers. The default answer was "No answer" and then the 10 possible answers were listed in alphabetical order.

Idioms for each section were taken from Street et al., a pilot study conducted using the American National Corpus to organize idioms. In that study, 4,500 sentences (68,915 tokens) were selected from the corpus from nonfiction, fiction, and spoken uses and classified into three categories: verb-noun phrase (VNP), prepositional phrase (PP), and subordinate clause (SC). Annotators either marked the sentences as idiomatic or not, and then the group of 4,500 sentences was broken down to 154 token idioms. Of these, 18% were PP, 79% were VNP, and 3% were SC. In an attempt to fit with these findings, my test has 30% PP and 70% VNP, with no SC, as only three of this type were included in the 154. The section had three idioms that are prepositional phrases and seven idioms that are verb-plus-noun phrase, randomly ordered.

Read (2004) suggests that for multiple-choice vocabulary tests,

the stem (i.e., the question) should be one or two simple, declarative sentences of 10-20 words. The questions for the present study's test were modeled after multiple-choice questions (i.e., although there were more options than four, all possible answers were provided) and so this suggestion was taken into account. Eight out of 10 sentence-level questions did not have subordinate clauses, all were declarative, and 5 out of 10 were simple sentences (i.e., no coordinating conjunction). Because the test was designed to be taken at nearly any level of proficiency, items were put through the vocabulary sorter from Tom Cobb's *Compleat Lexical Tutor* (www.lextutor.ca). Sentence-level questions were analyzed as 93.26% words from Nation's 1-1,000 most common English words, 2.59% second 1,000 words (5 types, 5 tokens), 2.59% off-list (this includes proper names). The cloze passage had 90.45% 1,000 most common words, 8.28% the second 1,000 words (only 8 types, but 13 tokens), 0% Academic Word List (AWL), and 1.27% off-list.

Procedure

The online version of the test was sent out via email to the researcher's contacts as well as to 12 teachers working at the American Language Institute at CSULB. The teachers forwarded the test to a class selected by the researcher for level (beginning students were deemed too low to participate) and to make sure that there was no overlap in students. The test was allowed to accrue participants for one month, at which point the test no longer accepted answers.

A second group of participants (17) performed a think-aloud protocol task while taking the test. They were first given instructions on think-aloud protocols using five grammar fill-in-the-blanks. The researcher modeled for the first question (i.e., pretended to be a student answering the questions), checked for understanding on the second question, and allowed the participant to continue without further instruction given that the participant seemed to understand the task. The participants then took the online test and voiced their thought process throughout the three sections of questions. As the participants voiced their reasoning aloud, the researcher tallied their responses into seven categories:

1. Guess;
2. Unknown/blank (i.e., no answer);
3. Grammar (i.e., participant used the grammar of the phrase or sentence);

4. Word association (i.e., the participant mentioned that certain words sounded alike or were associated, for example, "time" and "clock");
5. Context (i.e., participant used the context of the sentence or surrounding sentences);
6. Idiom known (i.e., participant already knew the idiom, but usually used context to place it in the right blank);
7. Other (including last answer left, and other responses that did not fit the preceding six categories).

After the think-aloud task, the participants were interviewed in person by the researcher and asked to elaborate on the Opinion Questions. All interactions were audio recorded.

Results

Research Question 1

A 4x4 factorial ANOVA was run to determine whether there was a significant difference among tests (matching, sentence, and paragraph). The assumptions for homogeneity and sphericity were not violated, so the main and interaction effects could be analyzed. Results for the differences between test type indicate that there was a significant difference between tests $p < .000$, $F\ (2, 72) = 12.809$. The effect size was large ($\eta^2 = .149$) and observed power was high (.997). There were no significant interaction effects. Post-hoc tests were run to determine which tests had scores that were significantly different. Results indicate that matching and sentence test scores were significantly different ($p < .000$) and matching and paragraph test scores were significantly different ($p = .001$), but that sentence and paragraph test scores were not significantly different ($p = 1.000$). Table 2 shows the descriptive statistics for the three test types.

Table 2
Descriptive Statistics for Test Type

	n	***Mean***	**SD**
Matching	85	4.318	2.518
Sentence	85	6.000	2.73
Paragraph	85	5.541	3.172

Research Question 2

The 4x4 factorial ANOVA also explored the impact of language level and native language on scores in the three types of tests given (matching, sentence, and paragraph). Subjects were divided into four groups based on native language (Arabic, Chinese, Romance, Other) and four groups based on English language level (high-beginner, intermediate, high-intermediate, and advanced). Results indicate that there is no significant main effect for native language, F (3, 73) = .250, p = .861, nor for interaction effect, F (5, 73) = 1.229, p = .304, but there was a significant main effect for language level, F (3, 73) = 6.411, p = .001. Effect size was large (η^2 = .209) and the observed power was high (.961). Post-hoc tests were run to determine which levels were significantly different from each other. Means and standard deviations for groups are indicated in Tables 3-6. Significant differences are summarized in Table 3. Tables 4-6 show the descriptive statistics for the English level group for the matching, sentence, and paragraph tests.

Table 3
Summary of Significant Differences

	High-beginner	***Intermediate***	***High-intermediate***	***Advanced***
High-beginner		1.000	**0.004**	**0.000**
Intermediate	1.000		**0.026**	**0.000**
High-intermediate	**0.004**	**0.026**		0.821
Advanced	**0.000**	**0.000**	0.821	

Table 4
Descriptive Statistics by Group: Matching Test

	n	***Mean***	**SD**
High-beginner	25	3.240	1.763
Intermediate	31	3.839	2.115
High-intermediate	18	4.778	2.487
Advanced	11	7.364	2.767

Table 5
Descriptive Statistics by Group: Sentence Test

	n	***Mean***	**SD**
High-beginner	25	5.160	2.703
Intermediate	31	5.161	1.899
High-intermediate	18	6.778	3.135
Advanced	11	9.000	1.612

Table 6
Descriptive Statistics by Group: Paragraph Test

	n	***Mean***	**SD**
High-beginner	25	4.400	2.858
Intermediate	31	4.323	2.427
High-intermediate	18	6.778	3.264
Advanced	11	9.546	0.934

Research Question 3

As a second and more exploratory aspect of the research, the think-aloud protocols were analyzed descriptively for frequency. Table 7 summarizes the percentage of usage for each method that participants used to help them figure out answers.

Table 7
Percentage of Usage of Reasoning Strategy

	Matching	***Sentence***	***Paragraph***	***Total***
Guess	24.85%	9.47%	8.88%	14.40%
Unknown/ blank	10.06%	6.51%	5.33%	7.30%
Grammar	2.37%	17.75%	14.79%	11.64%
Word association	56.80%	16.57%	6.51%	26.63%
Context	0.59%	39.96%	47.93%	29.38%
Idiom known	1.78%	8.28%	15.38%	8.48%
Other	3.55%	1.78%	1.18%	2.17%

Discussion

Research Questions 1 and 2

Providing context for vocabulary assessment is important for several reasons. Because teaching techniques often follow the communicative approach, testing with context is closer to the teaching methods used. This means that tests are more accurate reflections of what the students have learned as opposed to rote memorization. By that same token, context-based questions require deductive reasoning. Deductive reasoning is often used when students encounter new words in reading or speaking. A test that requires deductive reasoning can be a better predictor of how students will perform in reading and speaking. Because tests such as the TOEFL are mostly interested in predicting a student's success, it seems reasonable that test makers would want to get a more accurate picture of how a student would perform on a task of a similar type (i.e., no context is rote memorization whereas context involves reasoning, and communication in another language is also a cognitively complex activity).

Additionally, context questions are more accurate representations of a word's meaning. In matching tasks, students must attempt to match a word to its synonym. However, synonyms may show a less complete picture of a word's nature by leaving out collocations, approximate usage (i.e., how often a word is used), and of course the context in which a word is used. In terms of negative washback, students who study vocabulary by relying on synonyms are missing the complete knowledge of a word. By creating tests that require students to know the context in which a word is used, it is also possible that a positive washback could occur, which would influence students to a more complete understanding of vocabulary usage.

Results from the ANOVA differ from the findings from Uçkun's study (2008). In her study, there was no significant difference between scores on the matching and sentence-level questions, but there was a significant difference between scores on matching and paragraph and sentence and paragraph. The results of the present study indicate that the nonsignificant difference is between sentence and paragraph. This indicates that the addition of context is the factor that affects score for participants at all levels. Uçkun found that the matching tests resulted in the highest scores, whereas this study found that sentences had the highest scores. Matching had the lowest scores for the present study, but cloze was the lowest for Uçkun's study. These very different findings may be related to the difference in environments. In Uçkun's study, the participants were complete classrooms who had seen the vocabulary they were being tested on, and vocabulary was selected because it was appropriate for the proficiency level of the students be-

ing tested. In the present study, the test was designed specifically to not correspond to any classroom vocabulary or level. At core, the tests were fundamentally different, which could explain the difference in results. The present study's test was designed to be difficult for lower levels and easier for higher levels, but Uçkun's test was created to work for each specific level being tested, which could account for the higher overall scores for advanced speakers in the present study. It is unclear exactly why participants scored highest in matching on Uçkun's test and highest in sentence in the present study, but it is highly possible that the main factor was that participants had assuredly encountered the vocabulary before in Uçkun's test, but that they may have never seen the idioms in the present study's test. Additionally, the change from lexical items in Uçkun's study to idioms in the present study could account for differences, as idioms are known to have opaque meanings, although it should be noted that any single word other than onomatopoeia does not correlate to anything in real life other than the word as a symbol for the concept or thing.

Participants scored the lowest on matching test items, which supports the evidence that suggests that contextualized tests can be a more effective measure of a student's knowledge of a vocabulary word. As the scores for sentence and paragraph were not significantly different, it appears that the increase in context from zero is more important than the slight increase from sentence to paragraph. Because the scores were the highest for the sentence-level questions, it may be that overly contextualized test items may confuse students. It is also possible that the additional cognitive load of having to comprehend a paragraph as opposed to a sentence may have burdened the participants rather than aided them.

Studies have shown that learners with more advanced language skills are better able to manipulate the use of context when deciphering answers on vocabulary tests (Uçkun, 2008). The present study corroborates this finding as there was no significant difference between high-level speakers (high-intermediate and advanced). Low-level speakers (high-beginner and intermediate) also showed no significant difference. However, there was a significant difference between high-level speakers and low-level speakers. As Laufer (1987) points out, at least 95% of a passage must be understood before language learners can use the context to help them figure out new vocabulary items. If high-beginning and intermediate students are not understanding at least 95% of the text, then they would be unable to use the context. Additionally, if a reading passage is missing items (i.e., the vocabulary tested), deciphering meaning may be even harder. This is a pos-

sible explanation for why there was no significant difference between sentence-level items and paragraph-level items.

Research Question Three

The think-aloud protocol offers interesting insight into how participants reasoned their answers. A test of idioms is by nature different from traditional word testing because idioms usually cannot be "figured out" as they are not directly translatable or relatable to what they are said to mean; for this reason, this type of test offers a unique perspective on the strategies students use to rationalize vocabulary answers. In the first set of questions, the matching type, there was no context for the participant to use other than the fact that idioms are often phrases. As such, participants relied heavily on word association. For example, one idiom, *up-to-the-minute*, was often associated with the answer *all the time* because time deals with minutes. *Odd* was known to mean *strange*, so one participant chose for *to be at odds* the logical choice of *to meet unexpectedly*, as *unexpected* was associated with *strange*. Word associations are also related to collocations. With the idiom *have your share*, one participant associated *share* with *share ideas*, which he then extrapolated to *share information*, which led him to the answer *newest information*. *Up-to-the-minute* was also associated with the phrase *What's up?*, which asks for information. The participant chose the answer *newest information* because information was associated with *up.* Many participants, especially native speakers of Arabic, misread the word *dawn* to be *down*, not an unexpected result given the research that shows that Arabic speakers may have difficulty recognizing vowels in written forms of English words (Hayes-Harb, 2006). While word association was the most popular device that participants used to reason their responses, it should also be noted that matching had the highest percentage of mere guesses (24.85% compared to 9.47% and 8.88% on sentence and paragraph, respectively) and "no answer" responses (10.06% compared to 6.51% and 5.33% on sentence and paragraph, respectively). This suggests that participants may have been less confident with their reasoning ability given that there was no context to help them figure out the answers.

Confidence levels were recorded during the interview process. Participants were asked on a scale of 1 to 10, with 1 being all answers were wrong and 10 being all answers were correct, how confident they were in their results for each test. The average confidence level for matching tests was 5.07. The average of sentence tests was 6.71, and the average for paragraph tests was 6.36. The confidence levels correlate with the performance on the tests. Participants performed

slightly better on sentence than paragraph (.459 points higher), and their confidence level was also slightly higher (.35). The score difference between matching and sentence and matching and paragraph was 1.682 and 1.223, respectively. Participants were less confident in their matching scores, and the only participants who rated themselves higher in matching than sentence or paragraph also said that they preferred the matching test to the other two.

The most common strategy for the sentence-type questions was context (39.96%). This is not surprising, given that students are often taught to use the context to help them figure out the answers to unknown vocabulary encountered during reading. Some participants even used the word *context* during the think-aloud protocol and then afterward when interviewed about their responses. The second most common strategy was using grammar (17.75%). When participants used grammar, they seemed more sure of their answers, even if they were incorrect, as if there were something more concrete for them to rely on than mere definition.

The paragraph-type questions produced results similar to those of the sentence, except that the use of context was higher (47.93%), while the use of grammar was reduced (14.79%). This is possibly because the participants had more context to rely on and so therefore often chose context over grammar for a rationale.

The most popular strategy overall was context, which was used 29.38% for the total number of responses. As two of the tests were contextually based, it is not surprising that this strategy was so well employed. This, and the high percentages for both sentence and paragraph, shows that when the context is available, participants exploit it most often as compared to other strategies. When context is not available, word association is the next most common strategy, with 26.63% of the total responses. This means that participants choose to use other words to help them figure out the answers to questions on vocabulary tests. Rather than blending their vocabulary skills with grammar, the majority of participants chose to use known words to help them decide on the meanings of unknown words. While it has been pointed out that nonnative speakers do not perform well on tasks that require them to use context to decipher unknown vocabulary (Deighton, 1959; Laufer, 1987; Schatz & Baldwin, 1986), it does appear that they employ this strategy liberally, either as a consequence of what they have been taught or as a carry-over from deciphering in their own language.

On a cognitive level, the think-aloud protocol indicates that participants are aware of their test-taking strategies in terms of vocabulary and that they employ a variety of techniques to reason their answers.

Sasaki (2000) performed a similar test on cloze passages in which she had students give verbal reports of their test-taking strategies, similar to the present study's think-aloud protocol. She found that cloze tests are able to measuring higher-order thinking skills, a claim supported by others (Bachman, 1982, 1985; Chávez-Oller, Chihara, Weaver, & Oller, 1994). Considering that there was no significant difference between sentence-level questions and paragraph-level questions, it seems reasonable to suggest that the sentence-level questions may be tapping into higher-order thinking skills as well. While the lower-level thinking skills are classified as being at the clausal level (Sasaki, 2000), and a sentence is at that level, because of the lack of significant difference between these two types of tests, it is plausible to think that the sentence-level questions may be possibly requiring higher-order thinking, especially as many of the sentence-level questions actually comprised two short sentences (as recommended by Read, 2000).

Conclusion

This study looked at how context is manipulated during vocabulary-assessment tasks of idiomatic expressions in ESL students from the high-beginner to advanced levels. Three levels of context were used: matching (no context), sentence-level gap filling (medium context), and a rationale-delete cloze passage (high context). Tests were designed to test participants' knowledge of idiomatic expressions common in the English language but perhaps previously unseen by the participants. One group of participants was given a think-aloud protocol task while taking the test in order to determine how participants figured out the answers to the tests.

Results showed that there was a significant difference between no-context and contextualized test questions (both sentence and paragraph), which was contrary to the original study (Uçkun, 2008) and also to Qian (2008), which looked at vocabulary items on the TOEFL test. In the present study, there was no significant difference between contextualized tests, which would seem to suggest that context is what is needed, not the amount. Because the present study's findings differ from the prior studies, further research in the field of vocabulary assessment in terms of context is needed.

There are some possible limitations to the test. One limitation is that only 10 items per context type were included. This was mostly because of the increased likelihood of participant fatigue if the number of questions were increased. Additionally, because the test uses idioms instead of lexical items, degrees of transparency of these items may have been different. As such, it is possible that a collection of more transparent idioms was placed into any one context type. Although

the researcher tried to combat this by attempting to select idioms that are well known to native speakers, this does not completely address the chance that scores may have been affected by the degree of transparency. Future studies would do well to address these issues when creating tests of varying context.

While there are limitations to the study and while further research is needed because of the conflicting nature of the results in relation to prior studies, this study is important in that it is one of a handful of studies that empirically tests the idea of the supposed importance of context when creating vocabulary assessments. Because classroom teachers may be more concerned with negative washback when creating their tests, perhaps they should be more careful about how they design their tests in order to make them line up with teaching methods. But in the world of language-proficiency testing, efficiency can also play a more important role. Students have to have the time to answer more cognitively demanding questions (i.e., ones with context). Test makers need to consider ease of grading and how well a test actually tests a student's knowledge of all levels of a word, not just recognition (as is the case with matching). Because of the complex nature of vocabulary learning, vocabulary testing can also be a complex field. The present study recommends that test makers use a variety of techniques when testing vocabulary knowledge. As cloze passages do not appear to differ significantly from sentence-level questions, the researcher recommends that sentences are favored over cloze passages because sentence-level questions are simpler to create.

Author

Brenna Shepherd is a recent graduate of California State University, Long Beach with an MA in Linguistics, TESOL. She is a faculty member in the University of California, Irvine's Department of Academic English and hopes to continue her research in the field of vocabulary acquisition and assessment with new student populations.

References

Arnaud, P. J. L. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 133-145). London, England: Macmillan.

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*, 61-70.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535-556.

Chávez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W. Jr. (1994). When are cloze items sensitive to constraints across sentences? In J. W. Oller Jr. & J. Jonz (Eds.), *Cloze and coherence* (pp. 229-245). London, England: Associated University Press.

Cobb, T. (n.d.) *Compleat lexical tutor.* Retrieved from www.lextutor .ca.

Deighton, L. (1959). *Vocabulary development in the classroom*. New York, NY: Bureau of Publications, Teacher's College, Columbia University.

ETS. (2012). *About the TOEFL iBT Test.* Retrieved from http://www .ets.org/toefl/ibt/about/

Folse, K. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor: University of Michigan Press.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7*(5), 219-224.

Hayes-Harb, R. (2006). Native speakers of Arabic and ESL texts: Evidence for the transfer of written word identification processes. *TESOL Quarterly, 40*(2), 321-339.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range and frequency programs [Software]. Retrieved from http://www.victoria.ac .nz/lals/about/staff/paul-nation

Henning, G. (1991). A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items. *TOEFL Research Reports*, *35*. Princeton, NJ: Educational Testing Service.

Hughes, A. (1989). *Testing for language teachers*. Cambridge, England: Cambridge University Press.

Laufer, B. (1987). The lexical perspective of reading comprehension. *English Teachers' Journal (Israel)*, *35*, 58-67.

Laufer, B. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*(2), 202-226.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307-322.

Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size, and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Lewis, M. (1993). *The lexical approach*. Hove, England: Language Teaching.

Lewis, M. (1997). *Implementing the lexical approach.* Hove, England: Language Teaching.

Marks, C., Doctorow, M., & Wittrock, M. C. (1974). Word frequency and reading comprehension. *Journal of Educational Research, 67*(6), 259-262.

McQueen, J. 1996. Rasch scaling: How valid is it as the basis for context-referenced descriptors of test performance? In G. Wigglesworth & C. Elder (Eds.), *The language testing cycle: From inception to washback* (Series S, No. 13, pp. 137-187). Canberra, Australia: Applied Linguistics Association of Australia.

Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society: Papers from the annual meeting of the British Association of Applied Linguistics* (pp. 80-87). London, England: CILT.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research, 53*, 253-279.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233-253.

Nation, I. S. P. (1983). Teaching and testing vocabulary. *Guidelines, 5*(1), 12-25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge, England: Cambridge University Press.

Qian, D. (1998). *Depth of vocabulary knowledge: Assessing its role in adults' reading comprehension in English as a second language* (Unpublished doctoral dissertation). University of Toronto, Toronto, Canada.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*, 282-308.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513-536.

Qian, D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly, 5*(1), 1-19.

Read, J. (1989). *Towards a deeper assessment of vocabulary knowledge*. Paper presented at the 8th Congress of the International Association of Applied Linguistics, Sydney, New South Wales, Australia.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*, 355-371.

Read, J. (2000). *Assessing vocabulary*. Cambridge, England: Cambridge University Press.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24,* 146-161.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7*(2), 105-125.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing, 17*(1), 85-114.

Schatz, E., & Baldwin, R. (1986). Context clues are unreliable predictors of word meanings. *Reading Research Quarterly, 21*(4), 439-453.

Street, L., Michalov, N., Silverstein, R., Reynolds, M., Ruela, L., Flowers, F., … Felman, A. (2010). Like finding a needle in a haystack: Annotating the American National Corpus for idiomatic expressions. *Proceedings from LREC '10,* Valletta, Malta.

Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18*, 523-534.

Taylor, J. (1998). Syntactic constructions as prototype categories. In M. Tomasello (Ed.), *The new psychology of language* (pp. 177-202). Mahwah, NJ: Lawrence Erlbaum.

Uçkun, B. (2008). How does context contribute to EFL learners' assessment of vocabulary gain. *The Asian EFL Journal, 10*(2), 102-131.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *The Canadian Modern Language Review, 53*, 13-40.

Wilkins, D.A. (1972). *Linguistics in language teaching*. Cambridge, England: Cambridge University Press.

Wu, M. L, Adams, R. J., & Wilson, M. R. (1998). ACER ConQuest: Generalised item response modeling software [Software]. Camberwell, Victoria, Australia: ACER.