Routledge
Taylor & Francis Group

Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: http://www.tandfonline.com/loi/hlaq20

Aspects of Validity of a Test of Productive Vocabulary: Lex30

JoDee Walters

To cite this article: JoDee Walters (2012) Aspects of Validity of a Test of Productive Vocabulary: Lex30, Language Assessment Quarterly, 9:2, 172-185, DOI: 10.1080/15434303.2011.625579

To link to this article: http://dx.doi.org/10.1080/15434303.2011.625579

Language Assessment Quarterly, 9: 172–185, 2012

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2011.625579

Routledge
Taylor & Francis Group

Aspects of Validity of a Test of Productive Vocabulary: Lex30

JoDee Walters
Bilkent University

This study investigates aspects of validity of an alternative measure of productive vocabulary. Lex30, developed by Meara and Fitzpatrick, is a word association task that claims to give an indication of productive vocabulary knowledge. Previous studies of Lex30 have assessed test–retest reliability, performance against native speaker norms, concurrent validity, reliability of parallel forms, and ability to reflect improvements in vocabulary development. In addition, the issue of construct validity has been explored. The study described here replicates some of these investigations with a different population and extends the investigation of construct validity. By comparing the performance of second language (L2) learners at different proficiency levels, the ability of the test to distinguish between levels of proficiency is explored. Concurrent validity is explored by comparing L2 learners' performance on Lex30 with that of two other productive vocabulary tests. Finally, one aspect of construct validity is explored by assessing whether Lex30 measures productive vocabulary use or simply recall. The findings indicate that Lex30 is a reliable and valid measure of productive vocabulary knowledge, but whether it measures only recall, or whether it measures actual ability to use vocabulary meaningfully and appropriately, appears to depend on the proficiency level of the test taker.

BACKGROUND

Testing productive language skills has always been a difficult endeavor, and this situation is no different in the area of testing vocabulary, particularly when what is wanted is a way of measuring how much vocabulary is known by a language learner (i.e., breadth rather than depth). There are two methods available to vocabulary testers: checking the test taker's knowledge of selected words at different frequency bands, or eliciting as many words as possible from the test taker and analyzing those words for frequency. The former method is used in the Productive Vocabulary Levels Test (PVLT; Laufer & Nation, 1999), whereas the latter is employed in the Lexical Frequency Profile (Laufer & Nation, 1995) and in P_Lex (Meara & Bell, 2001).

In the PVLT, the test taker is asked to complete a word situated in a sentence that provides a certain amount of context for the target word. The beginning of the word is provided (typically two to four letters) in order to prevent the test taker from supplying a word that might fit the meaning of the text but is not the target word. It is important to remember that this test is considered

Correspondence should be sent to JoDee Walters, Monterey Institute of International Studies, Graduate School of Translation, Interpretation and Language Education/MA TESOL/TFA, 460 Pierce St., Monterey, CA 93940. E-mail: jwalters@miis.edu

to be a test of *controlled* productive vocabulary, because the test taker's response is necessarily restricted to one word—the target word. However, it is possible that the test taker might choose a different word to complete the sentence, with the further chance that this choice might be a less frequent word, possibly indicating a broader productive vocabulary than the test would reveal.

The Lexical Frequency Profile (LFP; Laufer & Nation, 1995) looks at vocabulary size (or rather richness, as an indicator of size) by eliciting many words from the test taker, through a free writing task, and then analyzing the frequency of the vocabulary used, with a tool such as Vocabprofile (Cobb, n.d.). The words used by the test taker are assigned to four different categories: the first 1,000 most frequent words in English (K1); the second 1,000 (K2); words that appear on the Academic Word List (Coxhead, 2000); and "off-list" words, those words not appearing on any of the other lists. The K1 level is further divided into function and content words, and content words are divided into the first and second 500. The Lexical Frequency Profile presents the percentage of words in the writing sample occurring at each level, and the free productive vocabulary behavior of writers can be compared to that of other writers, or to their own writing at different times or on different tasks. P_Lex (Meara & Bell, 2001) also seeks to describe the vocabulary produced by test takers by considering its frequency, but it uses different methodology. Rather than percentages of words at different frequency levels, a single score (a lambda score) is produced that corresponds to the proportion of infrequent words to frequent words. Meara and Bell (2001) claimed that this measure is more stable than the LFP with the shorter compositions likely to be produced by lower proficiency language learners.

Both the LFP and P_Lex have the advantage of not restricting the test taker in terms of what words he or she produces, because the words are produced in the context of a relatively free writing task (although the topic is assigned). However, these methods of estimating vocabulary size require that a great deal of text be elicited to generate a suitable amount of infrequent vocabulary. As Meara (2009) pointed out, any piece of writing (i.e., connected discourse) will necessarily contain a high proportion of high-frequency words, which contribute little to an understanding of a test taker's breadth of vocabulary knowledge. For example, the compositions (containing about 300 tokens) produced by participants in Laufer and Nation's (1995) study contained large proportions of K1 words ($M = 74$–87.5%; p. 316). It is also worth noting that free-writing tasks can be time-consuming; the participants in the two studies just mentioned were given 1 hr to produce their compositions.

Given the difficulties just described, in seeking a picture of productive vocabulary knowledge, it is not surprising that other options are being considered. One such alternative is Lex30, developed by Meara and Fitzpatrick (2000). This test is essentially a word association task, in which test takers are asked to write four associated words in response to a stimulus word. Test takers' responses are not constrained in any way, other than the request to "write four words which you think are related to it [the stimulus word]." The stimulus words are high-frequency words, which Meara and Fitzpatrick claimed makes the test appropriate for use with a wide range of ability levels. Each test taker's responses (up to 120 words) are subjected first to lemmatization, and then to a frequency analysis. Responses that fall in Level 0 (high-frequency function words, proper nouns, and numbers) and Level 1 (first 1,000 content words) receive zero points. The Lex30 score consists of the number of responses that fall outside of these two levels. The entire testing and scoring procedure can be carried out on computer, and a web version can be found at http://www.lognostics.co.uk/tools/Lex30/index.htm.

Meara and Fitzpatrick (2000) presented the results of a study that investigated the performance of an early version of the Lex30 test with a group of 46 adult English as a Foreign Language (EFL) learners from a variety of L1 backgrounds. The subjects ranged in proficiency from elementary to intermediate, as rated by their teachers. Scores on the Lex30 ranged from 1 to 60, with a mean score of 28.9. The average number of words produced was 91.6. The Yes/No Vocabulary Size Test (Meara & Jones, 1987), a test of receptive vocabulary size, was also administered, and the scores on this test correlated positively with the Lex30 scores (.841, $p < .01$). Investigation of the distribution of these two sets of scores indicated that, although for the most part receptive and productive vocabulary sizes are proportional (i.e., that productive vocabulary tends to be smaller than receptive vocabulary but that the larger the receptive vocabulary, the larger the productive vocabulary), the test would also be useful in identifying learners whose vocabulary development might be skewed toward one dimension or the other, so that remedial action might be taken. For example, the data revealed several participants whose productive vocabulary was much smaller than might be predicted by their receptive score, or vice versa (p. 27). The authors concluded by suggesting that Lex30 would be appropriate alongside other tests of vocabulary, given the speed and ease of administration and scoring, but they stated the need for exploration of the reliability and validity of the test.

This concern is addressed in Fitzpatrick and Meara (2004), who conducted a variety of studies to investigate the reliability and validity of the Lex30 test. In the first reliability study, test–retest reliability was examined. The test was administered to the same group of subjects (16 second language [L2] learners, lower intermediate to advanced level) twice, with a gap of 3 days between test administrations. The scores from the two administrations correlated strongly (.866, $p < .01$), and comparison of the words produced in both administrations revealed that all participants produced new words in the second administration, although the profile of those words remained essentially the same.

Fitzpatrick and Meara (2004) also looked at the validity of the test in terms of native speaker norms. In this study, the scores of 46 native speakers of English were compared with those of the 46 L2 learners in the study reported in Meara and Fitzpatrick (2000). The native speaker group scored significantly higher than the L2 group, but it was seen that there was some overlap between the scores of the two groups. Some L2 subjects scored higher than some L1 subjects, and only six L1 subjects produced higher scores than the highest scoring L2 subject. This overlap was explained by examining the characteristics of the highest scoring L2 subjects, who were found to have produced extremely high scores on the test of receptive vocabulary. Four of these subjects were Icelandic secondary school teachers of English, and the fifth subject was a very advanced German student of English. From this validity study, the authors concluded that the Lex30 test has some degree of validity, in that it works well in distinguishing between very proficient and less proficient users of the language.

Finally, Fitzpatrick and Meara looked at the concurrent validity of the Lex30 test, comparing its performance with that of the productive version of the Vocabulary Levels Test (PVLT) and of a translation test from first language (L1) to L2. The subjects were 55 Chinese English as a Second Language students, rated by their teachers as being at an intermediate to advanced level of proficiency. Although they were administered all five levels of the PLVT (2,000, 3,000, 5,000, University word list, and 10,000 levels), correct responses were almost all at the 2,000 and 3,000 levels. The translation test consisted of 60 Mandarin words, chosen from the 1,000, 2,000 and 3,000 levels of Nation's (as cited in Fitzpatrick & Meara, 2004) word lists, with

20 words from each list. The first letter of each target word was provided to ensure that the intended word was produced. The scores of the three tests were correlated, and substantial correlations were seen among all three tests (Lex30, PVLT, .504, $p < .01$; Lex30, translation, .651, $p > .01$; PVLT, translation, .843, $p < .01$), although the authors express surprise that the correlations are not as strong as expected. The strong correlation between the PVLT and the translation test was least surprising, given the fact that both tests were focused on the first 3,000 words of English. Fitzpatrick and Meara suggested that the relatively weaker correlations between Lex30 and the other two tests might be explained by the fact that the tests are measuring different aspects of vocabulary. Referring to Nation's description of the aspects of word knowledge, Fitzpatrick and Meara stated that although the PVLT addresses five of the eight productive aspects of word knowledge (knowledge of written form, grammatical position, collocations, appropriateness, and meaning), Lex30 measures only knowledge of written form, meaning, and associations, and the translation test measures only written form and meaning. In further discussion of the construct being measured by the test, Fitzpatrick and Meara raised the point that, although the Lex30 test appears to be measuring the recall dimension of productive vocabulary (Read, 2000), it gives no information regarding learners' ability to use that vocabulary. The authors conclude by reiterating that the Lex30 test is a useful test for providing information about one aspect (productive recall) of vocabulary knowledge and is appropriate for use alongside other tests of vocabulary knowledge.

Fitzpatrick and Clenton (2010) conducted further analysis of the reliability and validity of Lex30, exploring the reliability of parallel forms, its internal consistency, its ability to reflect improvements in vocabulary knowledge, and some aspects of the construct being measured. Parallel forms were found to correlate well (.692, $p < .01$), and the means of the parallel forms were not significantly different. A calculation of Cronbach's alpha produced a result of .866, indicating acceptable internal consistency. To determine whether the test would reflect vocabulary improvement over time, Lex30 was administered to the same group of L2 learners twice over an interval of 6 weeks, during which the learners participated in a language improvement class. The scores of the two test administrations were compared, and it was seen that the mean of the second test was significantly higher than that of the first administration. To determine whether Lex30 measures productive vocabulary in general, or only written productive vocabulary, as was suggested by Baba (2002), written and spoken forms of the test were compared. Although the means were not significantly different, the correlation between the two forms was low (.391, $p < .01$), leading the authors to question whether the vocabulary produced by the written form of the test would replicate that produced by a spoken form. Finally, to conclude their discussion of the construct validity of Lex30, Fitzpatrick and Clenton examined the theoretical bases of how vocabulary is elicited in the test, how that vocabulary is measured, and what it represents. They did not, however, address the question, raised by Fitzpatrick and Meara (2004), of whether the test measures only productive vocabulary *recall*, or whether it gives any information about test takers' ability to *use* the words they produce.

The purpose of the study to be described in this article is to further explore the validity of the Lex30, specifically its ability to distinguish between learners at different proficiency levels; its concurrent validity; and, to a limited extent, its construct validity. The study attempts to replicate some aspects of the experiments just reported but also expands the exploration into one specific aspect of construct validity by investigating the participants' ability to use the words produced in the Lex30 test. The research questions addressed by the study are as follows:

1. Does the test adequately distinguish among learners of different proficiency levels?
2. How do test scores compare with those of other tests that also claim to measure productive vocabulary?
3. To what extent are test takers able to *use* the words produced on the test?

THE STUDY

Participants and Setting

This study was conducted with 87 EFL learners in three different educational settings at universities in Turkey. All participants were from the same L1 background (Turkish). Thirty-two of the participants were studying in an MA/TEFL program at Bilkent University and had at least 2 years of English teaching experience in university language preparation programs. Twenty-five participants were in their 3rd year of an English teacher preparation degree program at Erciyes University. The remaining 30 participants were in their second semester of a 1-year English language preparation course at Hacettepe University. It was not possible to measure the language proficiency levels of the participants. However, their experience with the English language was used to categorize them into rough proficiency groups. Table 1 illustrates the rationale behind the categorization.

It should be noted that these are relative categorizations; based on their experience with the English language, the advanced group is more proficient than the intermediate group, which in turn is more proficient than the high-beginning group. Although there may be some variation within the groups, each group is believed to be distinct from the other two groups.

Instruments

Data were collected using four instruments: the Lex30 test, the PVLT, a translation test, and a sentence elicitation task.

TABLE 1
Proficiency Level Descriptions

Participant Group	*English Language Experience*	*Proposed Proficiency Level*
Bilkent University group (N = 32)	Currently studying in an English-medium MATEFL program; a minimum of 2 years of English language teaching experience.	Advanced
Erciyes University group (N = 25)	Completed 1-year English preparatory program at university; currently studying in an undergraduate level English Language Teaching program, 3rd year.	Intermediate
Hacettepe University group (N = 30)	Currently studying in a 1-year English preparatory program at university, in second semester.	High beginning

Note. MATEFL = Master of Arts program in teaching English as a foreign language.

Lex30. Even though a computerized version of the Lex30 test is available (http://www.lognostics.co.uk/tools/Lex30/index.htm), it was decided to use a paper–pencil version of the test, to enable the extraction of low-frequency words for further testing (not possible on the web version for a group of participants). The same stimulus words were used as are seen in the computerized version, presented on a single page with spaces to write four responses per stimulus. The instructions given on the test paper were the same as on the computerized version, as was the example. To score the test, the participants' responses were assembled into individual text files and lemmatized, according to the procedure described in Meara and Fitzpatrick (2000). In the previous studies of Lex30, previously described, the scoring was performed using the JACET list of basic words (Ishikawa et al., 2003). In the present study, for ease of analysis, the text files were uploaded into Vocabprofile (Cobb, n.d.) to analyze the frequency of the responses, which were identified as K1, K2, AWL, or off-list words. All words falling into the latter three categories were awarded 1 point, and the total represented the score on the Lex30 test. This change in the scoring procedure is likely to have produced slightly higher scores in the present study (P. Meara, personal communication, April 16, 2010).

PVLT. To avoid frustration on the part of the participants, especially the high-beginning participants, and to allow for time constraints, only the first two levels of this test were used, the 2,000- and 3,000-level tests. There was some concern that the use of only these two levels would make it difficult to distinguish between the advanced and intermediate groups; however, although 21 of 30 advanced-level participants scored 12 or more points on the 3,000-level test (two thirds of the available points), only two intermediate-level participants scored 12 or more at this level. Thus, it seemed that clear differences would be observed between these two groups. The tests were given on paper. To score the test, 1 point was awarded for each correct answer. An answer was considered to be correct if it closely resembled the intended target word; spelling and grammar mistakes were ignored. All participants received a total score representing the combined total of correct answers on both the 2,000- and 3,000-level tests.

Translation test. The translation test consisted of 60 words, chosen from the 1,000, 2,000, and 3,000 frequency levels (according to the Brown corpus; Kucera & Francis, 1967), 20 words from each level. Turkish equivalents of these words were established with the help of two Turkish native speakers, and a bilingual Turkish-English speaker was asked to translate them back into English. Necessary adjustments were then made, including replacing words that generated problematic translations or caused difficulty, and the test was prepared by presenting the Turkish words in the left column, and a blank with the first letter of the target word in the right column. One point was awarded for each correct translation (spelling mistakes were not penalized), and total scores reflected the combined number of correct translations from all three frequency levels. As with the PVLT, there was a concern that a ceiling effect would obscure differences between the two higher levels, but again, this was not observed.

Sentence elicitation task. In an attempt to determine whether test takers' responses on the Lex30 test represented merely productive recall, or also the ability to use the words, a sentence elicitation task was devised for each participant. Asking test takers to write a sentence to show that they know the meaning of a word is used as part of the Vocabulary Knowledge Scale, a measure of depth of vocabulary knowledge, developed by Wesche and Paribakht (1996). Indeed, the ability to write a sentence using the target word represents the highest category of word knowledge

on this scale. However, in their study of vocabulary acquisition using this scale, Paribakht and Wesche reported some difficulty in determining whether a word was known, if the word was used in a very general way. Read (2000) also cautioned that the ability to write an appropriate sentence using the target word may not always be an indication of knowledge of meaning. In this study, such problems have, it is hoped, been controlled for through the use of a five-band scoring rubric (described next). However, in the absence of another, efficient method to deal with many participants and many words, this method was chosen to provide at least a rough idea of ability to use target words.

Due to time constraints and consideration of the burden on the participants, it was decided to concentrate only on the lowest frequency words for the sentence elicitation task, rather than focusing on all words that contributed to the Lex30 score. Each participant's responses falling into Level 3 (AWL and off-list words, as defined by Meara and Fitzpatrick, 2000) were compiled into an individualized sentence elicitation test; the word was presented, and the participant was asked to use the word in a sentence. No further instructions were given, apart from encouraging them to write a sentence that showed "that you know what the word means." To score this test, a rubric was developed by the researcher to judge how well the use of the word in the sentence reflected ability to use the word. The rubric consisted of five bands, from 0 to 4, with Band 4 representing appropriate use of the word in a meaningful sentence. This level allowed for grammatical errors elsewhere in the sentence and for errors in tense and subject–verb agreement. Level 0 represented incomprehensible sentences, sentences in which the participant's grasp of the meaning was clearly wrong, or those in which the word was not used. After the rubric was developed, the sentences were scored by the researcher and another native speaker of English, who was also an EFL teacher; the sentences had previously been typed by an assistant for ease of scoring and to allow for blind scoring. About 15% of the sentences were scored working together, to ensure that both raters were using the rubric consistently, and the remainder of the sentences were scored independently. The level of agreement between the two raters was .964, and items for which scores differed were resolved by discussion, focusing on the appropriateness and accuracy of the use of the word in the sentence.

Procedure

For all groups, tests were administered in two sessions. In the first session, the Lex30 test, the PVLT, and the translation test were administered, and the sentence elicitation task was completed in the second session. Participants were given as much time as they needed for each test. There was a 4-day interval between sessions for the advanced and intermediate groups, and a 3-day interval for the high-beginning group. Not all participants took all tests, due to absence during one of the sessions.

RESULTS AND DISCUSSION

Ability to Distinguish Among Proficiency Levels

The instructions for Lex30 state that participants "will need about 15 minutes to complete this test." However, to ensure that speed of access was not a factor in the measurement of productive

vocabulary, participants were given as much time as they needed to complete the test. In all three groups, one third to one half of the participants completed the test in 15 min, with the remaining participants taking between 15 and 30 min. In spite of this, not all participants provided the maximum number of responses (120). Table 2 shows the mean number of words provided by each proficiency level and by the whole group.

Table 2 shows that participants at all levels provided a wide range of numbers of responses but that the mean number of words provided increases as proficiency level increases. The descriptive statistics for the Lex30 test can be seen in Table 3, for all three proficiency levels and for the group as a whole.

Table 3 shows that the means for the three groups appear to increase with proficiency level, and a one-way analysis of variance (ANOVA) confirms that the means for the three groups are significantly different, $F(2, 84) = 72.591$, $p < .001$, $\omega = .99$. Post hoc Scheffé tests reveal that the means for all groups are significantly different from each other ($p < .01$). It can also be seen that there is considerable overlap among the groups, with 19 advanced participants scoring lower than the highest scoring intermediate participant and 13 intermediate participants scoring lower than the highest scoring high beginning participant. Moreover, two advanced participants scored lower than the highest scoring high-beginning participant.

Thus, it can be seen that the Lex30 test appears to distinguish among proficiency levels, in that the means of the three proficiency groups are significantly different. However, the overlap in scores seen among the groups would indicate that individuals at the same proficiency level will vary considerably in terms of their Lex30 scores.

Concurrent Validity

To examine concurrent validity, two other tests, purporting to measure at least some aspect of productive vocabulary, were administered along with the Lex30 test: the PVLT and a translation

TABLE 2
Number of Words Provided on Lex30 Test

	No.	*Minimum*	*Maximum*	*M*	*SD*
High beginning	30	27	109	59.40	15.99
Intermediate	25	52	120	87.28	21.586
Advanced	32	50	120	110.19	18.045
Whole group	87	27	120	86.09	28.073

TABLE 3
Results of Lex30 Test

	No.	*Minimum*	*Maximum*	*M*	*SD*
High beginning	30	16	37	27.23	5.722
Intermediate	25	20	59	36.72	10.048
Advanced	32	28	77	55.84	11.706
Whole group	87	16	77	40.48	15.549

test. The descriptive statistics for the combined scores on the PVLT test (2,000 and 3,000 levels) are presented in Table 4, for all proficiency levels and for the group as a whole.

As with the Lex30 test, the means appear to increase with proficiency level, and a one-way ANOVA confirms that the differences among the groups are significant, $F(2, 84) = 175.600$, $p < .001$, $\omega = .99$. Post hoc comparisons reveal that significant differences exist between all proficiency levels ($p < .001$). Also similar to the Lex30 test, there is some degree of overlap among the groups, with 18 advanced-level participants scoring lower than the highest scoring intermediate-level participant, and eight high-beginner-level participants scoring higher than the lowest scoring intermediate-level participant.

Table 5 shows the descriptive statistics for the combined scores of the translation test (1,000, 2,000 and 3,000 levels), for all proficiency levels, and for the group as a whole. It can be seen that the groups' means again increase as the level of proficiency increases. A one-way ANOVA confirms a significant difference among the means, $F(2, 81) = 144.467$, $p < .001$, $\omega = .99$, and post hoc comparisons reveal that there are significant differences between each of the groups ($p < .001$). As with the Lex30 and PVLT, there is overlap among the groups on the translation test.

To explore concurrent validity, the scores from the three tests were correlated, and the results are shown in Table 6. It can be seen that there are strong, positive correlations among all three tests, with the strongest correlation between the PVLT and the translation test; a stronger correlation between the PVLT and translation test was also seen in Fitzpatrick and Meara (2004), who put it down to the fact that both of these tests focus on the first 3,000 most frequent words in English. The slight decrease in the strength of the correlation between the Lex30 test and the other two tests can also be explained with reference to frequency levels, as the unconstrained nature of the Lex30 test allows test takers to respond with words from beyond the 3,000 level.

The correlations presented in Fitzpatrick and Meara (2004) were relatively weaker than those shown here. This may be a result of the inclusion of a wider range of proficiency levels in the study

TABLE 4
Results of Productive Vocabulary Levels Test

	No.	*Minimum*	*Maximum*	*M*	*SD*
High beginning	30	3	19	9.53	3.803
Intermediate	25	12	30	21.56	5.148
Advanced	32	23	36	29.12	3.508
Whole group	87	3	36	20.20	9.305

TABLE 5
Results of Translation Test

	No.	*Minimum*	*Maximum*	*M*	*SD*
High beginning	27	17	32	25.00	3.893
Intermediate	25	18	47	36.12	6.399
Advanced	32	39	56	46.41	4.039
Whole group	84	17	56	36.46	10.177

TABLE 6
Correlations, Lex30, PVLT, and Translation Test

	PVLT	*Translation Test*
Lex30	.772 ($p < .001$)	.745 ($p < .001$)
PVLT		.936 ($p < .001$)

Note. PVLT = Productive Vocabulary Levels Test.

reported here, and thus a wider range of scores on the tests being compared. Fitzpatrick and Meara did not report the range of scores obtained by their participants, who are described as ranging from "intermediate level to advanced." The participants in the present study included students at a fairly low level of proficiency preparing to study at a university in a foreign language setting, in which only some of their classes would be offered in English; students at an intermediate level already studying at university with some of their classes in English; and teachers of English as a foreign language, some of whom could be considered to be very advanced. It is to be expected that a wider range of scores would produce a stronger correlation, and vice versa.

From the results of this study, it can be seen that the Lex30 test shows a high degree of concurrent validity with two other tests of productive vocabulary. Although it may be true, as Fitzpatrick and Meara speculated in their study, that the Lex30 measures a different aspect of vocabulary than either the PLVT or the translation test, it appears that performance on the Lex30 test can, to some extent, predict performance on the other tests, and vice versa, at least among a wide range of proficiency levels.

Recall Versus Use

In discussing the construct validity of the Lex30 test, Fitzpatrick and Meara suggested that perhaps the Lex30 was testing only a limited aspect of productive vocabulary knowledge, that of recall. A more in-depth and theoretical discussion of the construct validity of the Lex30 is given in Fitzpatrick and Clenton (2010), but the distinction between recall and use was not addressed. In an attempt to explore this particular aspect of the construct validity of the test, this study included a sentence elicitation task, to see whether test takers could also use the words they were able to recall in association with the stimulus word.

For the sentence elicitation task, it was decided to consider only those sentences scoring a 4 (the highest score) on the rubric as evidence of ability to use the low-frequency words provided in the Lex30 test. Twenty-six of the high beginning participants completed this task, along with 24 intermediate participants and 30 of the advanced participants. Test takers were required to write sentences for only those words falling in Level 3 (AWL and off-list words). This meant that some participants wrote more sentences than others. Figure 1 presents the range of number of words for which the students were required to write sentences. Figure 1 shows that the participants in the advanced group were, with some exceptions, required to write many more sentences than those in the other two groups. More than half of those completing the sentence elicitation task in the high-beginning group wrote between 11 and 15 sentences, whereas about two thirds of the advanced group wrote more than 26 sentences.

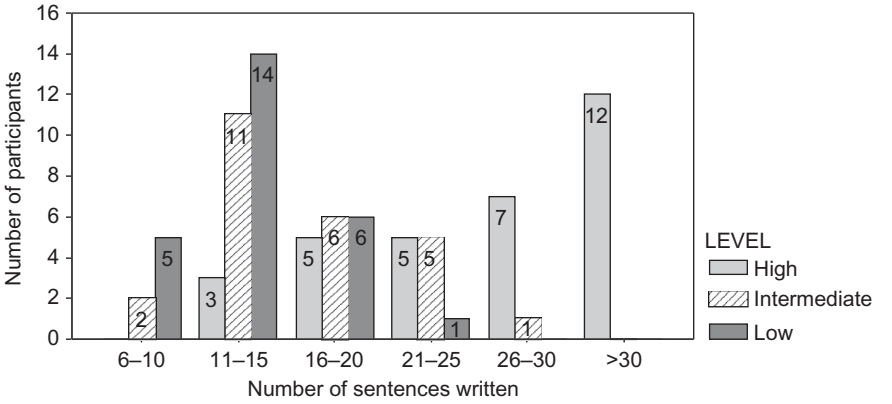

FIGURE 1 Number of sentences written, all groups.

TABLE 7
Results of Sentence Elicitation Task

	No.	*Minimum*	*Maximum*	*M*	*SD*
High beginning	26	.00	100.00	62.7603	27.80417
Intermediate	24	53.85	100.00	81.2868	12.34675
Advanced	30	74.29	100.00	88.8294	6.81710
Whole group	80	.00	100.00	78.0942	20.74436

Because the number of sentences varied among individual test takers, a percentage score for each participant was calculated by dividing the number of sentences scoring 4 by the number of words falling in Level 3 on the Lex30 test (i.e., all the words they were asked to write sentences for). Table 7 presents the descriptive statistics for these percentage scores.

Table 7 shows that the advanced and intermediate groups appear to be not only more successful than the high-beginning group in producing appropriate sentences using the lower frequency words they produced on the Lex30 test, but also more uniformly successful, with relatively more narrow ranges of scores and smaller standard deviations. The high-beginning group, on the other hand, included at least one student who produced no appropriate sentences and at least one student all of whose sentences were appropriate. In addition, the mean for this group is more than 25 percentage points lower than that of the advanced group. A one-way ANOVA revealed a significant difference among the scores, $F(2, 77) = 15.628$, $p < .001$, $\omega = .94$, and post hoc comparisons showed that the differences between the high-beginning group and the two higher groups are significant ($p < .01$). There is no significant difference between the intermediate and advanced groups.

It is clear that, in this study, higher demands were placed on some participants than on others. If fatigue had been a factor, negatively affecting the participants' ability to write appropriate sentences, one would expect to observe a negative relationship between the number of sentences written and the number of appropriate sentences written. However, no such relationship was

observed. In fact, the high-beginning participant who wrote the most sentences (21) scored well above the group mean (76% vs. 63%).

The figures just presented refer only to words that might be considered at the extreme end of the "known" continuum. If words that are partially known are included in the analysis, by also including sentences which scored a 3 on the rubric, a more favorable view of productive ability emerges. These sentences showed a good grasp of the meaning of the word, but there were problems with such aspects as prepositions or voice, or the student wrote an appropriate sentence using a different part of speech than that written on their Lex30 test. When these sentences are also considered, the means increase for each group, to 91% for the advanced group, to 87% for the intermediate group, and to 69% for the high-beginning group. For the group overall, the mean increases to 83%. This more lenient view of productive knowledge can also be seen in the scoring of the PVLT, in which spelling and grammatical mistakes are not penalized.

From these results, and keeping in mind the potential problems with this method of determining ability to use target words (previously mentioned; Read, 2000; Wesche & Paribakht, 1996), it appears that the Lex30 is perhaps a more valid measure of productive vocabulary *use* at higher proficiency levels than at lower levels. For an average of 91% of the lower frequency words that advanced learners recorded as associates of the stimulus words, they were able to demonstrate acceptable productive knowledge, in the form of written sentences using the target words, and intermediate learners were able to produce acceptable sentences for 87% of the words. In contrast, high-beginning students were only able to demonstrate such knowledge for an average of 69% of the lower frequency words they recorded as associates. It may be that, for the lower level participants, the task of writing a sentence was a considerable strain on their productive capabilities. However, an effort was made in the scoring criteria to discount problems in sentence construction and to focus only on the appropriate use of the word in the sentence. Thus, it appears that although the Lex30 test may be a valid test of productive vocabulary *use* for higher proficiency students, it is more valid as a test of productive vocabulary *recall* at the lower levels.

Although high-beginning participants had some difficulty producing appropriate sentences for roughly 30% of the low-frequency words they produced on the Lex30 test, it should not be assumed that they had no knowledge of those words. Roughly 13% of these participants' sentences received a score of 2 on the rubric, and nearly 6% received a score of 1, indicating some grasp of meaning. Only 11% received a score of 0. Thus, it can be concluded that most of these words were known to the participants (recalled), in spite of their inability to produce appropriate sentences using the words.

CONCLUSION

This study has attempted to explore the validity of a test of productive vocabulary knowledge, Lex30. It has been seen that, with these participants, the Lex30 is able to distinguish among proficiency groups. In contrast to the significantly different means among the proficiency groups, considerable overlap in scores was seen among the groups. However, similar overlaps were seen with the other two tests—the PVLT and the translation test. Such overlaps reflect the fact that there is considerable variation in vocabulary knowledge in learners at the same level. It is important to remember that the purpose of such tests is not to predict proficiency level but to predict productive vocabulary knowledge.

It has also been seen that the Lex30 shows good concurrent validity with two other tests of productive vocabulary knowledge. In addition, one aspect of the construct validity of the test has been explored by examining what learners are able to do with the words the Lex30 test would suggest are indicative of a certain breadth of productive vocabulary. It has been seen that, in this respect, the Lex30 test appears to behave somewhat differently at different proficiency levels, acting as a valid test of productive vocabulary recall at lower proficiency levels, but as a valid test of productive vocabulary use at higher proficiency levels.

However, some difficulty in interpreting Lex30 scores can be anticipated. One area of difficulty arises from the way points are awarded. Because all words outside of the most frequent 1,000 content words are awarded the same point value, regardless of their relative frequency, it would be possible for a score of 30 to represent either 30 words entirely from the 1,001 to 2,000 frequency range, or 30 words entirely from the 2,001 to 3,000 frequency range. Of course, it is unlikely that a participant would produce *only* words from a single frequency range, but the point remains that a single Lex30 score might represent a variety of vocabulary profiles. It may be that reporting a lexical frequency profile of the lower frequency words, rather than a single score, might make the results of the Lex30 more useful in comparing individual learners.

Another difficulty concerns what the scores actually indicate. Unlike the PVLT, which provides an estimate of the size of the test taker's productive vocabulary size in terms of number of words known (extrapolated from the number of correct answers), a score obtained from the Lex30 test does not correspond to an estimate of vocabulary size. Thus, Lex30 cannot be a replacement for a test like the PVLT. In its present form, it seems most useful as a way of comparing individuals in terms of their breadth of vocabulary knowledge. Given the strong correlations between the PVLT and Lex30, it seems fair to say that the higher the Lex30 score, the larger the productive vocabulary size. However, there is still the question of what a particular score *means* in terms of vocabulary knowledge. Perhaps with more extensive testing with participants at well-documented proficiency levels, typical Lex30 score ranges might come to be associated with particular groups of learners. Alternatively, given the spread of scores within proficiency levels seen in this study, it might be more useful to associate Lex30 score ranges with vocabulary sizes determined via the PVLT.

Although the study reported here was conducted with relatively small numbers, the findings show that the Lex30 test can be a useful and valid test of productive vocabulary knowledge, under certain circumstances. The test, particularly the computerized version, is easy to administer and score, taking less time than either the PVLT or the Lexical Frequency profile. It allows test takers to demonstrate the breadth of their vocabulary knowledge without constraint in a short time and allows for comparison among learners. If some attention is given by the test developers to clarifying how test scores are to be interpreted, it may be that Lex30 will indeed become "a robust enough measuring tool to fill an important gap in the battery of tests currently available" (Fitzpatrick & Meara, 2004, p. 72).

REFERENCES

Baba, K. (2002). Test review: Lex 30. *Language Testing Update*, *32*, 68–71.

Cobb, T. (n.d.). Web Vocabprofile, an adaptation of Heatley & Nation's (1994) *Range*. Retrieved from http://www.lextutor.ca/vp/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary. *Language Testing*, *27*, 537–554. doi:10.1177/0265532209354771

Fitzpatrick, T., & Meara, P. (2004). Exploring the validity of a test of productive vocabulary. *VIAL: Vigo International Journal of Applied Linguistics*, *1*, 55–73.

Ishikawa, S., Uemura, T., Kaneda, M., Shmizu, S., Sugimori, N., Tono, Y., . . . Murata, M. (2003). *JACET 8000: JACET list of 8000 basic words*. Tokyo, Japan: JACET.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

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

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*, 33–51.

Meara, P. M. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Herndon, VA: John Benjamins.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*, 5–19.

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*, 19–30.

Meara, P., & Jones, G. (1987, September). *Vocabulary size as a placement indicator*. Paper presented at the Annual Meeting of the British Association for Applied Linguistics, Nottingham, UK.

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

Wesche, M., & Paribakht, T. S. (1996). Assessing L2 vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review*, *53*, 13–40.