# Short-Term Study Abroad and Lex30: a Replication Study

Thomas H. Caton

（Received on November 22, 2017）

## Background

With short-term Study Abroad (SA) programmes measuring changes in language ability has been problematical. There are few tests available that are sensitive enough to cover short periods of language learning and the degree of language improvement (or otherwise) is difficult to measure (Drake 1997). An assessment method which focuses on a particular aspect of language and knowledge is more likely to reveal subtle changes taking place over a short duration than more general testing techniques.

It has been suggested that methods of lexical analysis might be sensitive enough to pick up changes in the overall language competence of short term SA programme participants. Fitzpatrick and Clenton (2010) conducted analysis of the reliability and validity of Lex30, a productive vocabulary size test which elicits samples of vocabulary using word association. They examined the reliability of parallel test forms, internal consistency and the ability to measure improvements in vocabulary knowledge. In order to find out if the test could detect vocabulary improvement over short time periods it was administered to the same group of learners at the beginning and at the end of a six-week language improvement course. The pre-course and post-course scores were compared and it was seen that the mean of the second test was significantly higher than the first. A similar improvement was found in a more recent study. Caton (2015) examined Japanese short-term SA participants' productive vocabulary size again using Lex30 and found the same improvements between pre and post programme values. Although this was only a small single case study on a group of 19 students, it suggested that further detailed examination could reveal more interesting results.

This replication study will explore some of the findings of Fitzpatrick's (2003) thesis research. In particular it will examine the results of the longitudinal study she carried out with SA participants to see how alternative methods of scoring the Lex30 can give very different results. The study will also look at a similar study carried out by Fitzpatrick and Clenton (2010) which tried to measure similar changes in productive vocabulary knowledge over a short period. Finally, a replication experiment, at least in part, conducted by the author in early 2016 on three groups of Japanese students participating in SA programmes will be described and the findings compared with previous two sets of results.

## Fitzpatrick and Clenton's (2010) paper

Fitzpatrick and Clenton's paper assesses many aspects of the performance of the Lex30 vocabulary test. They build on previous findings (Meara and Fitzpatrick 2000, Fitzpatrick and Meara 2004) and take a further look at the test's reliability and construct validity. The aim of their research is not to argue for, or against, the validity of the test per se but to thoroughly explore its potential and to identify its limitations. The overall usefulness of the test is considered by structuring the paper around a series of issues previously raised by other researchers (Baba 2002, Jiménez Catalán and Moreno Espinosa 2005). The authors also consider how the Lex30 test fits in with Bachman and Palmer's 'usefulness' formula (1996:18).

The paper describes the Lex30 test's ability to elicit lexically rich text in an economical way comparing it favourably with other productive knowledge measures including controlled productive knowledge tests (Laufer and Nation 1999) and the Lexical Frequency Profile

Reprint Request: Thomas H. Caton Nakamura Gakuen University Junior College, the Division of Career Development
Befu 5-7-1, Jonan-ku, Fukuoka, Japan 814-0198, E-mail：tomcaton@nakamura-u.ac.jp

(Laufer and Nation 1995). It describes the test as being able to elicit a wide range of vocabulary from different conceptual fields with using a single word association stimulus. It is argued that the careful selection of these 'cue' words minimizes the effort needed for their activation and maximizes the range of potential responses. Three main experimental areas are covered. Firstly, the reliability of the test is looked at using a test-retest study, a parallel test forms experiment and an internal consistency measure. Secondly, the paper looks at the test's construct validity trying to determine whether the test reflects vocabulary improvement over time by administering it to the same group of L2 learners over an interval of 6 weeks during which the learners participated in a language improvement class. I shall return to this particular experiment in due course as it forms the focus for this particular replication study. Finally, to conclude their discussion on the construct validity of the test, Fitzpatrick and Clenton examine some of the background theoretical bases of how vocabulary is elicited and measured.

The main question to focus on is whether the test reflects improvement in learners' vocabulary knowledge (Fitzpatrick and Clenton 2010:543). Prompted by comments from Baba (2002) calling for more evidence for the validity of Lex30 and following observations from Bachman (1990) about gathering such information by comparing learners of different language proficiencies it was decided to compare learner data with data from the same learners after a language learning intervention period when it might be reasonably expected that language proficiency might have improved.

## Fitzpatrick and Clenton's (2010) study

The longitudinal experiment was conducted using Lex30 to obtain criterion-related evidence on the validity of the test. It is designed to detect an vocabulary knowledge increase over a six-week period with a group of 40 L1 Japanese pre-intermediate students attending English improvement classes. The Lex30 test was administered on two occasions: at the beginning and at the end of the six-week 'language intervention' period. The list of responses to the 30 cue words was lemmatized according to strict criteria (from Bauer and Nation 1993 as described in Meara and Fitzpatrick' s 2000 paper). One point was awarded for every low-frequency word produced, with 'low-frequency' being defined as not being in the 1000 most frequently occurring English words. For this study the JACET 8000 word list (JACET 2003) was used.

The descriptive statistics are shown in Table 1. The difference between the two means (24 at test time one and 29 at test time two) was found to be significant ($t=4.825$, $p<0.0001$). The two sets of scores correlated at 0.809 ($p<0.01$). The authors conclude that the increase in scores shown here between test time one and test time two is evidence that the Lex30 test is capable of detecting a change, in this case an improvement, in learners' productive vocabulary ability. It was suggested that this significant improvement in scores was also not likely to be due to the practice effect of simply have done an identical test a second time.

## Fitzpatrick 2003 Thesis

Further building on her work with Paul Meara on the introduction of the Lex30 productive vocabulary test (Meara and Fitzpatrick 2000) Fitzpatrick explores how productive vocabulary can be elicited and measured by using word association techniques and word frequency lists. She describes Lex30 as a test which uses a word association technique to allow subjects to produce a small corpus of words which is representative of their total productive lexicon. The absence of predetermined target words and narrow context constraints encourage subjects to elicit content words across a wider range of frequency bands than might otherwise be the case.

Much of her research looks at the development

**Table 1. Fitzpatrick and Clenton 2010: Longitudinal study score data**

| | N | Min. | Max. | Mean | SD |
|---|---|---|---|---|---|
| Test time 1 | 40 | 9 | 42 | 24 | 8.514 |
| Test time 2 | 40 | 7 | 48 | 29 | 9.084 |

process, paying particular attention to cue or stimulus word selection by using the Edinburgh Associative Thesaurus (Kiss et al 1973 mentioned in Fitzpatrick 2003:115), a database of word association norms, listing response words, and the frequency with which they occur, for 8,400 stimulus words. Then the process of lemmatization is examined using the formal set of criteria from Bauer and Nation's "Word Family" lists (1993). Once the test has been shown to work relatively smoothly in practice, Fitzpatrick takes us through several more important stages looking at score consistency, native speaker comparisons and longitudinal studies generally looking at reliability and validity of the Lex30 test, and concludes that the test has significant potential as a measurement tool. She does caution us to be aware of concerns about its accuracy and sensitivity and this will become particularly evident when we take a look at longitudinal test validation studies.

## Fitzpatrick 2003 Longitudinal Study

The purpose of this study is to see if Lex30 can distinguish between non-native speakers in a useful way. It looks at two groups of students studying in Britain and tests each individual at the beginning and end of their study period during which time it is expected their language proficiency will undergo some kind of change. As well as taking the Lex30 test subjects also took the receptive EVST to compare any changes in their productive performance with changes in their receptive lexicon. The two study groups differ in that first was in Britain for a period of 4-weeks while the second, although in the country for a year-long exchange programme, was tested before and after a 5-month period. For this replication study and for the purposes of a comparison with the experiment carried out by Fitzpatrick and Clenton (2010) I shall only look at the first group.

The author looked 19 L1 Japanese undergraduate students participating in a 4-week intensive English language course at a university in the UK. Their age was between 19 and 23. During the course students received a minimum of three hours English language instruction per day while staying with local English-speaking host families. The subjects took the computer version of the Lex30 test on day one and were also tested 24 days later during the last week of their programme. The test required them to type in as many responses as possible (up to a maximum of 4) for each cue word provided. The Lex30 scores were calculated according to a percentage method (Fitzpatrick 2003:148-151). This means that each participant's Lex30 score represents the number of infrequent words they produce as a percentage of the total number of words produced. The subjects all took the Lex30 and the EVST together at both test times. For the purposes of this replication study I am only interested in the results of the Lex30 test.

The difference in the mean Lex30 scores between test time one and two was not significant. In other words, the Lex30 scores remained relatively stable over the 4-week period. The t value was: t=1.29 p=.213

The individual Lex30 performances at test times one and two show a significant correlation 0.636 ($p<.01$) between them. On the scatter graph (Fitzpatrick 2003:189) we can see the majority of subjects are placed above the line, indicating that they scored higher on test one than at test time two. A summary of the results suggests that the number of infrequent words in the subjects' productive lexicons has not increased over the study period or perhaps that the Lex30 test is not sensitive or sophisticated enough to pick up any increases over a short 4-week period.

## Towards a replication experiment

Both studies that have been described purport to measure a similar construct: detecting changes in

**Table 2. Fitzpatrick (2003): Group one: longitudinal study score data**

| | N | Mean | SD |
|---|---|---|---|
| Test time 1 | 19 | 22 | 6.97 |
| Test time 2 | 19 | 20 | 6.78 |

the productive vocabulary performance of students attending short-term SA programmes. The results from each are very different and this cannot be easily explained. A real difference in test performance is likely to be a factor but other influences may be at work, too. Differences in scoring procedures, test protocols and even learning environment may also play an important role. The following replication experiment will try to follow these earlier longitudinal studies and take account of some of these influences. First, it will essentially ask the same research question: can a longitudinal study with a similar number of L2 participants using Lex30 detect similar changes in productive vocabulary knowledge and second: can the way in which the Lex30 is administered and scored have any influence on the final outcome?

## Methodology

The participants consisted of 38 female Japanese students aged between 18 and 21 years old attending three separate courses at two universities in Fukuoka, Japan. Table 3. shows the background profile of the students and pre-test and post-test times.

In order to obtain a sample of reasonable size for the experiment three separate groups were used. All students spent 17 days in total in their respective study abroad countries staying with host families and undertaking a programme offering a similar educational experience. The pre test was conducted eight days before departure for students going to Vancouver, three days before for students going to Canterbury and five days before for students going to Hawaii. The post test for the Vancouver and Canterbury students was carried out at the language school just before their return to Japan while Hawaii students completed their test within three days after their return. The pre test and post test times are shown in Table 3. The test was administered by three university staff members who conducted both orientation classes and accompanied students on their programmes. Students were given a time limit of 15 minutes to complete the test on each occasion.

After the test was completed it was processed according to protocol laid down in Meara and Fitzpatrick (2000). All responses were individually lemmatized so that inflectional suffixes (plural forms, past tenses, comparatives) and frequent regular derivational affixes (-able, -ly) were counted as base forms of these words. The full criteria used by Meara and Fitzpatrick corresponds to levels 2 and 3 of Bauer and Nation's 'word families' (Bauer and Nation 1993). Other protocols followed including discounting proper names, numbers, Japanese words and acronyms from the corpus created by each student. For a full list of protocols covered please see Appendix A. The JACET 8000 wordlist was used to select the 1000 most frequently occurring words. Each word in a test subject's individual corpus that did not appear on the list of 1000 most frequently occurring words was awarded one point. This total number of infrequently occurring words was used to calculate a final raw Lex30 score.

## Results

The descriptive statistics are shown in Table 4. The difference between the two means (18.24 at test time one and 26.16 at test time two) was found to be significant ($t=6.8854$, $p<0.001$). The two sets of scores correlated at 0.747 ($p<0.001$).

The increase in scores shown here between test time one and test time two seems to be evidence that the Lex30 test is capable of detecting a change, in this case an improvement, in learners' productive

**Table 3. Replication 2016: Study group participants**

| N=38 | University and department | Programme location | PreTest | Post test |
|---|---|---|---|---|
| 17 | Nakamura: Career Development | Vancouver, Canada | 8 days | 0 days |
| 19 | Nakamura: Nutrition | Canterbury, UK | 3 days | 0 days |
| 2 | Fukuoka University: Law | Hawaii, USA | 5 days | 3 days |

**Table 4. Replication 2016: Longitudinal study score raw data**

| | N | Min. | Max. | Mean | SD |
|---|---|---|---|---|---|
| Test time 1 | 38 | 5 | 39 | 18.24 | 8.159 |
| Test time 2 | 38 | 9 | 50 | 26.16 | 10.666 |

vocabulary ability. This significant increase in scores can be interpreted as being a result of the participants attending an English study abroad programme. It can be concluded that, in this case, Lex30 seems to be sensitive to improvements in learners' language ability. The individual Lex30 performances at test times one and two are illustrated on the scatter graph in Figure 1. The graph indicates the relationship between the scores at the two test times. The line on the graph is where subjects who scored the same at both test times would be plotted. The great majority of subjects are placed below the line showing they scored higher at test time two than test one.

## Discussion

A comparison of this replication experiment with Fitzpatrick and Clenton's findings shows many similarities. With almost the same number of participating students (38 and 40) both the minimum and maximum scores are very close for both the pre and post tests. For example the maximum score in the replication post test was 50 compared to 48 for Fitzpatrick and Clenton while the minumum score was 7 compared to 9. The spread of scores were also similar as shown by the figures for standard deviation. In both studies the difference between pre and post test means was significant although the t value was slightly higher with the replication. The correlation of test time one and test time two scores is also close with figures of 0.809 ($p<0.0001$) for Fitzpatrick and Clenton's study and 0.747 ($p<0.0001$) for the replication.

In order to make a meaningful comparison with Fitzpatrick's 2003 longitudinal study it was necessary to convert the raw scores (infrequently occurring words from JACET 1000+) gained with the replication study into a percentage form (described in Fitzpatrick 2003:148). Very different scores were obtained with the percentage scores compared with raw Lex30 scores. In both the replication study (see Table 5.) and Fitzpatrick 2003 the number of infrequent words produced by most participants increased from to test time one to test time two. However, the total number of responses increased over the same period at a much

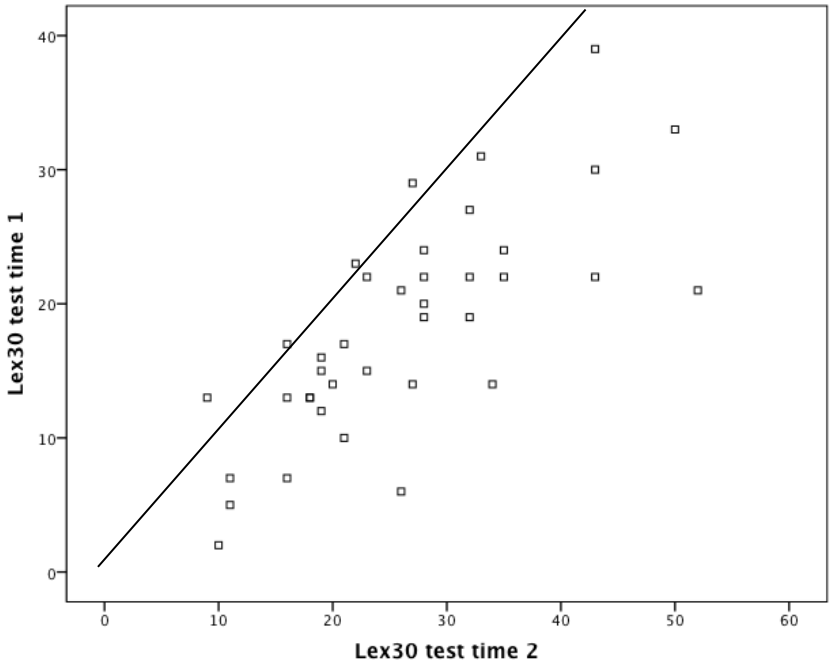


**Figure 1. Lex30 Raw Scores – 2016 Replication study**

Table 5. Replication study 2016: Longitudinal study percentage score data

| | N | Min. | Max. | Mean | SD |
|---|---|---|---|---|---|
| Test time 1 | 38 | 13 | 60 | 39 | 8.580 |
| Test time 2 | 38 | 14 | 49 | 35 | 7.680 |

greater rate. The result of this is that the percentage of infrequent words as a proportion of the total number of words produced by each participant actually decreased.

The difference in the mean Lex30 scores between test time one and two was not significant. As was the case with Fitzpatrick 2003 Lex30 scores remained relatively stable over the 17 day study abroad period. The t value was: t = 2.079.29 p=.045
The individual Lex30 performances at test times one and two show a slight correlation 0.406 (p<.406) between them.

The individual Lex30 performances at test times one and two are illustrated on the scatter graph Fig. 2. This graph should be compared with Fig.7.2 (P.189) describing the results of Fitzpatrick's original experiment. The graph indicates that there was a relationship between the scores at the two test times, and in fact there is a low correlation of 0.139 p< 0.406 This compared to 0.636 (p<0.1) in the original data (P.188). The line on the graph is where the subjects who scored the same at both test times would be plotted. It can be seen that the majority of subjects are placed above the line, indicating that they scored higher at test time one than at test time two.
There were a number of factors in the design and administration of the Lex30 test that might have an influence on the eventual outcome. I shall deal with each of these issues in turn.

## Participants

Given that female-only participant groups were chosen, the results of the current study might not relate to populations that do not share similar characteristics. Another consideration is that instead of three groups studying different courses at two universities perhaps it would better to have one homogenous group which can be better controlled. There was some variation in the arrangements for test administration during the experiment because the groups were meeting and preparing for their overseas trips at different times. However, I felt that having a sample size of 38 students

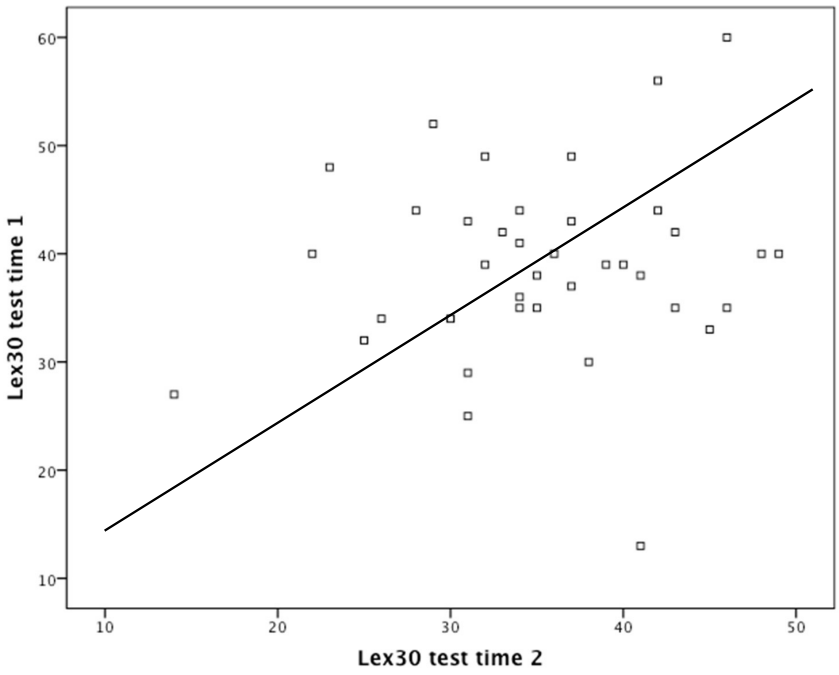


Figure. 2. Lex30 Percentage Scores – 2016 Replication study

would allow me to have more confidence in my results and was close to the 40 subjects participating in Fitzpatrick and Clenton's study.

## Participation period

This was a particular concern at the beginning but impossible to control for practical reasons. The length of short-term SA programmes has gradually shortened in recent years due to academic, economic and employment reasons. The 17 days spent by subjects on their short-term programme was considerably shorter than the four weeks and six weeks of the two studies that my experiment tried to replicate. There is evidence that the longer students spend studying abroad then the greater their gains in language proficiency. Milton and Meara (1995) found that SA students' vocabularies grew four times as fast compared to at-home learners and Llanes and Muñoz (2009) also correlated fluency gains with length of stay. Conversely, others argue that length of stay is less important that quality and quantity of contact with the target language while abroad (Bardovi-Harlig & Bastos, 2011). On balance I felt that the period of 17 days, although short, was still of sufficient length for the Lex30 to detect some sort of change.

## Timing of pre and post tests

Care was taken with the timing of both pre and post tests. If the pre tests are carried out too early then perhaps students would have further uncontrolled opportunities to increase their vocabulary proficiency before their departure and if post tests are delayed for too long after the return then the chance of vocabulary knowledge attrition would increase. I noted that with Fitzpatrick's 2003 study her subjects received their initial pre test only after they had spent the first weekend with their host family. During this two-day period, although short, there was some opportunity for latent vocabulary to be reactivated. Meara (2005) discussed evidence for the spontaneous reactivation of vocabulary knowledge and looked at data suggesting that his test subject's active vocabulary more than tripled in size over the course of just two days. His results should be treated with caution as he warned that they were not conclusive but it does seem that exposure to a L2 environment can quickly to encourage the reactivation of more frequently occurring words in particular. He also found that high frequency words were more likely more likely to be encountered at first upon initial exposure than low frequency. Although Meara failed to find conclusive evidence he suggested, "that we need to be aware that rapid and extensive vocabulary reactivation as a result of environmental input needs to be taken more seriously" (2005:279).

## Lemmatisation Procedure

I referred briefly to the lemmatization procedure earlier in the paper and criteria used by Meara and Fitzpatrick (2000) which corresponded to levels 2 and 3 of Bauer and Nation's 'word families' (Bauer and Nation 1993). I tried to award a mark for each infrequent vocabulary item and give, 'credit at every possible opportunity' (Meara and Fitzpatrick 2000:26) but processing each test subject's word corpus was sometimes problematical. Clenton (2005:53) gives some examples from a Japanese context including the use of Katakana (Japanese syllabic writing primarily used for words of foreign origin) in some of the responses and the use of loan words which, in practice, are used very differently in Japanese. Jiménez Catalán and Moreno Espinosa (2005) also looked at similar issues with Spanish students. I tried to overcome some of these difficulties by compiling a list of protocols (see Appendix A) and attempting to be as consistent as possible with their application.

## Percentage score v. Raw score

The last issue to be considered is the method of scoring. In earlier pilot testing Fitzpatrick (2003) used the raw Lex30 score as a basis to measure performance. She soon noticed, however, that there was a much greater variation in the number of words produced in response to Lex30 than there had been with earlier versions of the test and she became concerned that Lex30 was tending to emphasize corpus size over quality (in terms of high-frequency words) the corpus contained. She felt that any performance score calculated from the test task should be as independent as possible and should not allow the number of words produced to affect measurements of the quality of words. As a result

the raw Lex30 score was recalculated in terms of the number of infrequent words as a percentage of the total number of words produced therefore reflecting the proportion of infrequent words in each corpus. In both Fitzpatrick's (2003) study and in the replication there was a tendency to produce far more words at test time two than at test time one which might be, as expected, associated with language study between the two test times during both longitudinal experiments. In terms of the mean total number of words produced, Fitzpatrick (2003) study group increased from 73 to 94, a rise of 21 while the replication study group increased from 48 to 74, a rise of 26. This suggests that there was rise in vocabulary production or fluency within both groups. The mean number of infrequent words produced by all subjects also increased from 17 at test time one to 20 at test time two (Fitzpatrick (2003) and from 18 to 26 (replication). In both studies, figures for both the total number of words produced and the raw Lex30 scores increased but because the total number of words increased by considerably more the percentage score fell in both cases.

## Conclusion

The replication experiment tried to reproduce the results of two previous longitudinal experiments with mixed success. Comparisons with Clenton and Fitzpatrick (2010) were encouraging. Using a similar number of participants, similar test administration and protocols and most importantly the same Lex30 scoring system using a raw count of infrequently occurring words, the results obtained were roughly equivalent. With Fitzpatrick's 2003 study we are forced to reconsider what exactly it is we are measuring. Although it is perhaps illogical to have data suggesting declining levels of productive vocabulary over the course of an intensive SA programme, it does encourage us to start asking further questions. Is having a raw Lex30 score sufficient to make a judgment of subject's improvement in productive vocabulary knowledge over the course of a short-term SA programme or should account be taken of the influence of the total size of the corpus produced? Perhaps further exploration into how data from the Lex30 is processed will help us formulate new and better-balanced marking schemes for the future.

## Main References

Fitzpatrick, T. and Clenton, J. (2010) The challenge of validation: Assessing the performance of a test of productive vocabulary *LanguageTesting* 27(4) 537–554

Fitzpatrick, T. (2003) *Eliciting and measuring productive vocabulary using word association techniques and frequency bands* Unpublished PhD Thesis University of Swansea

## Other References

Baba, K. (2002) Lex30: Review *Language Testing Update* 32 68-71

Bachman, L. (1990) *Fundamental considerations in language testing* Oxford: Oxford University Press

Bachman, L., and Palmer, A. (1996). *Language Testing in Practice: Designing and developing useful language tests*. Oxford: Oxford University Press

Bardovi-Harlig, K., and Bastos, M. (2011) Proficiency, length of stay, and intensity of interaction and the acquisition of conventional expressions in L2 pragmatics. *Intercultural Pragmatics 8*, 347-384.

Caton, T. (2015) *The unexpected benefits of short term study abroad*. Paper presented at Japan Association of College English Teachers (JACET) Kyushu-Okinawa Chapter, Seinan University, Fukuoka, Japan, April 2015

Clenton, J. (2008) Addressing Baba: determining whether Lex30 is a reliable and valid testing measure *Studies in Language and Culture* 34 157-168

Clenton, J. (2005) Why Lex30 may not be an improved method of assessing productive vocabulary in an L2. *Studies in Language and Culture*, 31. pp. 47-59

Drake, D. (1997) Integrating study abroad students into the university community *The Language Teacher* 21 (11) 7-13

Fitzpatrick, T. and Meara, P (2004). Exploring the validity of a test of productive vocabulary. *Vigo International Journal of Applied Linguistics*, 1, 55-74.

Fitzpatrick, T. (2012) Tracking the changes: vocabulary acquisition in the study abroad context *The Language Learning Journal* 40 (1) 81-98

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

Jiménez, Rosa M. and Soraya Moreno (2005) Using Lex30 to Measure the L2 Productive Vocabulary of Spanish Primary Learners of EFL *VIAL Vigo International Journal of Applied Linguistics* 2: 27-44.

Llanes, A., and Muñoz, C. (2009) A short stay abroad: Does it make a difference? *System* 37 353–365

Laufer, B., and Nation, P. (1995) Vocabulary size and use: lexical richness in L2 written pro-duction. *Applied Linguistics* 16(3) 307–322

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

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

Meara, P. (2005) Reactivating a dormant vocabulary *EUROSLA Yearbook Vol 5* 269–280

Milton, J., and Meara, P. (1995) How periods abroad affect vocabulary growth in a foreign language *ITL Review of Applied Linguistics 107/108* 17-34

Porte, G (Ed) (2012) *Replication Research in Applied Linguistics* Cambridge University Press Cambridge

Schmitt, N. (2010) Researching vocabulary: *A vocabulary research manual*. Palgrave Macmillan

Walters, J. (2012): Aspects of validity of a test of productive vocabulary: Lex30 *Language Assessment Quarterly* 9:2 172-185

APPENDIX A : Lex30 Scoring protocols

This is a list of protocols followed when scoring the Lex30 test. Words from the JACET 1000 list were used. Every answer which occurred on this list was awarded '0' points. Answers which occurred which were not on this list were awarded '1' point. The number of misspelled (but acceptable) words and discounted words were also noted.

## Unacceptable

1. No proper nouns to be counted: *Japan, Canada, McDonalds, Kentucky*
2. No numbers
3. No acronyms to be counted: *USA DVD PC CM MC*
4. Japanese words – even those which appear to be an approximation of English words. Eg: *anime* unless they satisfy condition (5) below.
5. Prompt words (used as word association responses): A problem described in Jimenez Catalan and Moreno Espinosa (2005) P.41. However prompt words are acceptable as long as they satisfy condition (3) below.
6. Where two words are written for a single entry only one word will be noted. If one or both words are from L2+ category a maximum of one point will be credited. For example: Pot - *hot water* (counted as one word – credited with no points as both words are 1K level)
   Potato – *dietary fiber* (counted as one word – credited with one point even though both words are from L2+ category).

## Acceptable

1. Misspelled but still recognizable words, although JC and ME (2005) argue against this saying that there should be greater score weighting for correctly spelled answers.
2. Each response to the test was lemmatized so that:
   (i) Responses with an inflectional suffix (plural forms, past tenses, comparatives)
   (ii) Frequent, regular derivational affixes (-able, -ly )
   were counted as base-forms of these words. These criteria correspond to levels 2 and 3 of Bauer and Nation's "Word Families" (Bauer and Nation 1993)
3. The same answers written down for different prompt words as long as there is some kind of semantic relationship: furniture – *bed* and rest - *bed* seat – *movie* television – *movie* furniture – *sofa* seat – *sofa*
4. Some credit is given when a two part 'phrasal verb' is written instead of a single answer. This occasionally happened when students forgot about test instructions. Phrasal verbs were counted as one word. For example Hold – *take place* (scores 1 point) or Habit – *get up* (scores 1 point)
5. Words which are on the 'List of English words of Japanese origin' from Wikipedia. Example: *sushi* but not *Kanji* or *hiragana*