# Measuring the productive vocabulary of secondary school CLIL students: Is Lex30 a valid test for low-level school learners?——

*Rafael Alejo González*
Universidad de Extremadura
ralejo@unex.es

*Ana Mª Piquer Píriz*
Universidad de Extremadura
anapiriz@unex.es

## Abstract

Since it was issued (Meara and Fitzpatrick, 2000), Lex30 has been validated as an adequate instrument to measure L2 learners' productive vocabulary, mostly, in studies with university students (Fitzpatrick and Clenton 2010) but it has been also used with young learners in foreign language contexts (Jiménez Catalán and Moreno Espinosa, 2005; Moreno Espinosa, 2009; 2010). The study reported in this paper focuses on assessing the validity and reliability of Lex30 to measure the productive vocabulary of two groups of secondary school students (N=48) following a CLIL (Content and Language Integrated Learning) programme by analysing: 1) its reliability, 2) whether it correlates with general language proficiency, 3) if it measures vocabulary growth over long periods of time and 4) if it is sensitive to the possible effect of the context of learning on the productive vocabulary of the learner. The results suggest that Lex30 could be an appropriate test to be used with secondary school learners but they also seem to indicate that, especially in specific educational contexts such as CLIL, Lex30 scores should be interpreted with caution.

**Keywords:** Productive vocabulary, Lex30, validity, reliability, secondary school learners

## Resumen

Desde su lanzamiento (Meara and Fitzpatrick, 2000), Lex30 ha sido validado como un instrumento adecuado para medir el vocabulario productivo de aprendices de una segunda lengua, principalmente, en estudios con universitarios (Fitzpatrick and Clenton, 2010); pero, también se ha utilizado con alumnos de educación primaria en contextos de aprendizaje de lenguas extranjeras (Jiménez Catalán and Moreno Espinosa, 2005; Moreno Espinosa, 2009; 2010). El estudio que se presenta a continuación se

centra en analizar la validez y fiabilidad de Lex30 para medir el vocabulario productivo de dos grupos de alumnos de enseñanza secundaria (N=48) dentro de un programa AICLE (Aprendizaje Integrado de Contenidos y Lenguas Extranjeras), analizando: 1) su fiabilidad, 2) su correlación con otras medidas lingüísticas, 3) si mide el crecimiento del vocabulario en períodos largos de tiempo y 4) si es sensible al posible efecto del contexto de aprendizaje en el vocabulario productivo del alumno. Los resultados indican que Lex30 puede ser un test adecuado para alumnos de secundaria pero también parecen indicar que, en particular, en contextos específicos como el AICLE, las puntuaciones obtenidas deben interpretarse con precaución.

**Palabras clave:** Vocabulario productivo, Lex30, validez, fiabilidad, alumnos de enseñanza secundaria

## 1. Introduction

Ever since Lex30 was issued fifteen years ago (Meara and Fitzpatrick, 2000), its validity as an adequate instrument to measure productive vocabulary has been confirmed by the different studies published (Baba, 2002; Fitzpatrick and Meara, 2004; Fitzpatrick, 2007; Fitzpatrick and Clenton, 2010; Walters, 2012). As is well-known, validating a test is an on-going process whose main goal is to find evidence showing (or challenging) that it can adequately be used to measure the construct it purports to gauge. In order words, the validity of a test, i.e. showing that it actually measures what it purports to measure, is not established once and for all and, as suggested by recent literature on testing (see Kane, 2011, for a review), it resembles more a discussion where arguments and counterarguments are provided.

The best account of a validity argument in favour of Lex30 as an appropriate instrument to measure productive vocabulary of EFL learners has been given by Fitzpatrick and Clenton (2010) in their state-of-the-art article on this test. In this article, the authors claim that the test "produces reliable scores, reflects improvement in language knowledge, [and] produces scores which are comparable to those of similar tests" (2010:16). At the same time, in light of the relative weakness of some of the correlations found, the researchers also acknowledge that "Lex30 should not be used for […] any high-stakes testing in an educational context (to make absolute decisions about placement, proficiency, etc." (2010:549).

However, these general conclusions on the validity of Lex30 are not definitive and may be complemented by investigations further exploring the validation analysis with other populations (cf. for example Walters, 2012). Thus, given that university students make up the bulk of the learners analysed in validating the test (Meara

and Fitzpatrick, 2000; Fitzpatrick and Meara, 2004; Fitzpatrick, 2007; Fitzpatrick and Clenton, 2010; Walters, 2012), it would be valuable to establish whether Lex30 can also provide accurate data about the productive vocabulary of young and adolescent learners, with peculiarities of their own in language assessment processes (cf. McKay, 2006). Besides, in close relation to this, it would also be interesting to explore the ability of the test to register vocabulary growth over longer periods of time (years rather than weeks), a criterion-related validity feature, not analysed by the Lex30 literature up to now, which is particularly relevant for school settings (cf. chapters 2 and 4 in Milton, 2009 for an analysis of receptive vocabulary).

This task becomes all the more important as different projects analysing the vocabulary of young learners have used the test (cf. Jiménez Catalán and Moreno Espinosa, 2005; Moreno Espinosa, 2009), and given that few instruments analysing productive vocabulary adapt so well to the characteristics of young learners: 1) It is simple (it consists of 30 cue words that belong to the 1k band so these are words that even low-level learners are very likely to be familiar with); 2) It is accompanied by simple instructions and a clear example; 3) It does not need much preparation and it can be easily administered in a classroom context either on-line or in its pen-and-paper version; 4) it does not require much time to be completed, a maximum of 15 minutes; 5) it is context-independent, unlike other tests measuring L2 learners productive vocabulary, Lex30 elicits isolated words, and 6) as opposed to what happens with free productive written or spoken production, the task is standard and the same for all testees, and no task effects are to be expected; therefore, the scores are more susceptible to being compared.

This analysis is also necessary because some other features of the test may make it less appropriate to be used with younger populations in school settings. As stated by Milton (2009:143), a "potential fatal flaw of the test" may be that it requires learners to "willingly engage with the purpose of the exercise and […] not to maximise their scores rather than reflect their knowledge", an approach that may not be easily assumed in school contexts where the closest association to this type of test is an exam.

All in all, an analysis of the usefulness of Lex30 to determine the productive vocabulary of younger populations of learners in educational settings would greatly help to establish whether the noted advantages of the test do actually correspond with similar advantages in terms of psychometric properties (reliability and criterion validity). In this article, we analyse the validity of Lex30 as a test to measure the productive vocabulary size of two groups of secondary school CLIL students in Extremadura and explore the suitability of the test to be used with populations of teenage students learning English in instructional contexts.

## 2. Background : main features

### *2.1. Lex30 as a productive vocabulary test*

Lex30 constitutes one of the available tests affording to tap the construct termed productive vocabulary (Telchrow, 1982; Laufer, 1998). Other options measuring the same construct are the Productive Vocabulary Levels Test (PVLT, Laufer and Nation 1999), and instruments such as the Lexical Frequency Profile (LPF, Laufer and Nation, 1995) or the P-Lex (Meara and Bell, 2001), which focus on vocabulary use instead of vocabulary knowledge. However, the research on productive vocabulary is far from attaining the same consensus reached around the Vocabulary Levels Test (VLT, Nation, 1990, Schmitt, Schmitt and Clapham, 2001), which is the closest we have to a "standard test in vocabulary" (Meara, 1996:38 cited in Schmitt, 2010:197) and the tests of productive vocabulary can only be said to assess "broadly similar constructs" (Fitzpatrick and Clenton, 2010: 545) since they deal with different aspects of vocabulary knowledge (cf. Fitzpatrick and Meara, 2004).

Lex30 can also be described as a 'form recall' test (cf. Laufer and Goldstein, 2004 cit. in Schmitt, 2010), which is a way of indicating that the form and meaning relationship constituting a word is being accessed in the test via the meaning, cued by the prompt word, and the test taker task is asked to provide the form. In other words, the test measures the ability of learners to produce words when they are prompted to do so but not their ability to use the word (cf. Fitzpatrick and Meara, 2004).

### *2.2. Elicitation and scoring*

The elicitation procedure is an association task where the learner is asked to provide a maximum of 4 different words (3 in the initial version of the test) that could be freely associated with each word from a list of 30 cue words giving a maximum of 120 answers for the whole test. In the design of the test, the cue words selected were carefully chosen among the words included in the Edinburgh Association Thesaurus (EAT, 1973), a corpus of native-speaker associations, complying with all of the following features (cf. Fitzpatrick and Clenton, 2010:541): a) they belonged to the 1k frequency band; b) they had no strong primary responses in EAT, i.e. they were provided by less than 25% of test takers and; c) they elicited a high proportion of infrequent words, which meant that half of the responses should not belong to the 1k band. It is important to note here that, although the test is mostly, if not solely, used with non-native speakers, the criteria used in its design make reference to native speaker language as, to the best of our knowledge, there is no established word association thesaurus for non-native speakers. The suggestion made by more recent

research (cf. Fitzpatrick et al., 2015) is that norms lists need to be developed to match, among other things, the age and language status (native vs. non-native speakers) of the specific population studied.

The scoring procedure consists in counting the number of words outside the 1K band produced as associations by the learners. From this count, proper nouns, names, repeated words and words used as cues are excluded and the final tally is carried out using frequency bands taken from the JACET 8000 list (2003), a development included in later versions. The resulting scores give values ranging from 21.3 to 30 for non-native speakers and of 44 for native speakers (cf. Meara and Fitzpatrick, 2004), although they are sometimes higher (high beginning 27.23; intermediate 36.72; and advanced 55.84, cf. Wolter, 2012) when the scoring list is different (Vocabprofile Cobb, n.d.). In general, as noted by Fitzpatrick and Clenton (2012:548), the frequency lists used "are compiled from adult language", which may "potentially undermine[s] the theoretical validity of using this test with young learners".

### *2.3. Reliability and validity*

The seminal paper introducing Lex30 (Meara and Fitzpatrick, 2000) did not include an in-depth analysis of its reliability and validity. Since then, research has been published providing evidence of the compliance of the test with the general requirements of language assessment (summarised in Table 1 below).

**Table 1**: summary of studies on reliability and validity of Lex30

| Source | Quality explored | Method | N | Subject / L1 Background | Findings |
|---|---|---|---|---|---|
| Fitzpatrick and Meara (2004) | Reliability | Test-retest | 16 | Different L1 backgrounds | No difference in scores at T1 and T2. |
| Fitzpatrick and Clenton (2010) | Reliability | Test-retest | 103 | - | No difference in scores at T1 and T2. |
| Fitzpatrick and Clenton (2010) | Reliability | Parallel forms | 40 | Medical students Japanese | No difference in scores between parallel forms. |

| | | | | | |
|---|---|---|---|---|---|
| Fitzpatrick and Clenton (2010) | Reliability | Internal consistency | 35 | - | Cronbach alpha=0.866 |
| Fitzpatrick and Meara (2004) | Concurrent validity | Comparing NS and NNS scores | 46 native 46 non-native | Adults Different L1 backgrounds | Difference between NS and NNS. Some group overlap |
| Fitzpatrick and Meara (2004); Fitzpatrick (2007) | Concurrent validity | Comparing performance on different tests | 55 | Adults Chinese | Moderate correlations with PVLT and translation test |
| Walters (2012) | Concurrent validity | Comparing performance on different tests | 87 | University students Turkish | Correlations with PVLT and translation test |
| Walters (2012) | Concurrent validity | Comparing performance of groups with different levels | 87 | University students Turkish | Distinguishes between levels. |
| Fitzpatrick and Clenton (2010) | Construct validity | Test-retest (written-spoken) | 40 | University students Japanese, Chinese, Korean | No difference between written and spoken versions, but low correlation |
| Fitzpatrick and Clenton (2010) | Construct Validity | Test-retest | 40 | University students Japanese | Significant differences in scores over a period of 6 weeks |

The first element, reliability, which also contributes to the validity (Bachman, 1990), has been studied in different ways: 1) using a test-retest procedure (Fitzpatrick and Meara, 2004), 2) exploring the performance of parallel forms and 3) analysing its

internal consistency (cf. Fitzpatrick and Clenton, 2010). In general terms, the results in this area show that a considerable amount of words are not repeated between different sittings of the test and that the correlations between test-retest and the different versions range between moderate and moderately high values (ranging between 0.692 and 0.842).

The concurrent validity of the test has also been approached using different methods (cf. table 1), including comparing the performance of different groups (native vs. non-native speakers –cf. Fitzpatrick and Meara (2004) or beginning-intermediate-advanced learners, cf. Walters (2012)– and comparing the scores with other measures of productive vocabulary (cf. Fitzpatrick and Meara, 2004; Walters, 2012). In both cases, the results of the analysis are positive with some limitations with regard to the certain degree of overlap in the scores obtained by the different groups (native and non-native speakers in the first study and the different language levels in the second) and to the low correlations scores obtained.

Construct validity has also attracted some attention in trying to demonstrate the usefulness of the test to measure spoken productive vocabulary and vocabulary growth (cf. Fitzpatrick and Clenton, 2010). Like most vocabulary tests, Lex30 is based on a frequency construct, although a major caveat is that the results obtained from the test cannot easily be extrapolated to the actual vocabulary size learners have at their disposal (cf. Meara and Olmos, 2010), which makes it less apt to measure vocabulary growth in terms that are readily understandable. This area requires further exploration as the only study that refers to the vocabulary gain made by students in a period of 6 weeks (cf. Fitzpatick and Clenton, 2010), and the progress of learners usually needs more time to develop.

Finally, as stated in our introduction, and as can be seen from table 1, most of the subjects used in these reliability and validity studies are university students, with different L1 backgrounds. Young learners do not figure prominently in them, although, as we will see in our next section, Lex30 has frequently been used with this population of learners.

### *2.4. Lex30 in educational contexts*

The reluctance observed above to include young learners in the validity and reliability analysis of the test contrasts with the readiness to use it as a tool to measure productive vocabulary in educational contexts (cf. Fiztpatrick, 2007). For instance, Jiménez Catalán and Moreno Espinosa (2005) and Moreno Espinosa (2009 & 2010) used Lex30 with young learners in school contexts. In the first of these studies, we find that the L2 learners tested (4th graders) achieve very low scores as more than

50% of testees had scores between 1 and 10 and that the correlation values between Lex30 and PVLT were also very low at the two levels analysed (r=0.369 at the 1K level; and r=0.293 at the 2k level). The second study (Moreno Espinosa, 2009) was designed to establish whether two groups of Primary School learners (6th graders), one CLIL and one non-CLIL, differed in their vocabulary size as measured by Lex30 as a consequence of the learning context. Although the author shows that CLIL students have a statistically significant larger productive vocabulary than the non-CLIL group and the findings confirm a previous analysis of a writing task, the mean scores (19.51 vs. 17.56 respectively) can be considered slightly higher than the ones obtained in the previous study by Jiménez Catalán and Moreno Espinosa (2005) taking into account that their populations are only separated by two academic years of instruction. The results offered by Moreno Espinosa (2010), a study on the differences between the productive vocabulary of boys and girls, are also in line with the scores of her CLIL study, showing a progression in the productive vocabulary of the different grades analysed (11.41 for 4th grade, 14.54 for 5th grade and 17.4 for 6th grade).

Given the range of scoring values obtained by adult learners reviewed above (between 20 and 30), some of the results obtained by these authors appear to contradict Spanish proficiency levels in English at school (cf. *First European Survey on Language Competences,* 2012) and also some of the findings for this country in terms of passive vocabulary. As shown by Canga (2013), in his review of the research on the receptive vocabulary of Secondary School learners, Spanish learners at this educational stage, for whom the curriculum does not establish a target, do not typically go beyond the 1,000 word level, regardless of the number of hours of instruction received. One possible explanation may lie in the fact that they are analysing CLIL learners, who have been generally shown to perform better in this country (Ruiz de Zarobe, 2009; Lorenzo, Casal and Moore, 2009).

In any case, given the demonstrated lag of productive vocabulary with respect to passive vocabulary (cf. Takala, 1984, Laufer, 2005, Fan, 2000 or Laufer and Paribakht, 1998, cit. in Schmitt, 2010), no study has been undertaken to show whether for Lex30 to work properly a threshold level is necessary. In general, convergent validity studies relating Lex30 to other language measures is an area where further research is needed.

## 3. Research questions

From what we have seen in Section 2, even though the validity of Lex30 has been analysed extensively, there are still areas that need further exploring with respect to its usefulness as an adequate instrument to measure the productive vocabulary of young learners. In this article, we will try to address the following ones:

Does Lex30 elicit a reliable and representative sample of secondary school learners' vocabulary? (reliability)

Does vocabulary size as measured by Lex30 correlate with other measures of language proficiency and academic attainment at school? (criterion validity)

Does the test adequately measure vocabulary growth over longer periods, which are more typical of school processes? (criterion validity)

To what extent is native frequency adequate to score the test when the input accessed by the students is skewed by the context? (construct validity)

## 4. The Study

### *4.1. Context*

The present investigation is the result of a wider research project where we studied the affective and linguistic outcomes of CLIL in 9 schools in different geographical locations of the Extremadura region (SW Spain). In this paper, we focus on two groups of students enrolled in two different schools situated in an urban (school A) and a rural area (school B), respectively. Like the rest of Spanish autonomous regions, the Extremaduran autonomous government has certain legislative powers over education and in the case of CLIL it has a lot of leeway to set up specific policies. Some of the main characteristics of CLIL in Extremadura are shared with other regions, especially Andalusia, as they mostly follow the recommendations of the European Commission. Thus, 'Bilingual Section Projects', which is the name given to CLIL programmes in this region, are voluntary programmes for both Primary and Secondary School students and usually consist in the teaching of 2 or 3 content subjects through the medium of a foreign language (mostly English, some French and very little Portuguese). They usually involve the provision of an additional foreign language, typically French or Portuguese in the case of the majority English CLIL programmes (for a detailed account of the main features and the development of CLIL in Extremadura, see Alejo and Piquer, 2010: 228-233).

### *4.2. Participants*

The total number of students taking part in the study was 48: 27 students in school A (11 females and 16 males) and 21 in school B (12 females and 9 males). We gathered the data at two different moments, Times 1 and 2 (T1 and T2), which correspond to November 2009 and May 2011, the years in which both groups of students were

doing their 3rd and 4th year of Spanish Compulsory Secondary Education (*Enseñanza Secundaria Obligatoria*). Both groups, thus, received the instruction corresponding to nearly two full academic years that consists of approximately 140 hours of EFL instruction and about 70 hours of CLIL. In both schools, the CLIL programme included three subjects. In the case of school A those subjects were Social & Natural Sciences, Arts, and Technology. In school B, the subjects were Maths, Physical Education and Music. The students were between 14 and 15 years of age at T1 and 15 and 16 at T2.

### *4.3. Instruments*

As stated above, Lex30 (Meara and Fitzpatrick, 2000; Fitzpatrick and Meara, 2004) was chosen as the test to measure these learners' productive vocabulary because it fitted the constraints of time, and the age of the students made it preferable to other existing tests (e.g. Productive Vocabulary Levels test). Although productive vocabulary and the validity of Lex30 to measure it in secondary school learners are the main focus of this paper, further data from the research project can shed some light on some of the research questions posed for this study, particularly, on RQ2, which gauged the correlation between the results of productive vocabulary and other language variables: in our case, grammatical knowledge and receptive vocabulary. The grammatical knowledge of the students was analysed by adapting some of the tests used by *Dialang* (Alderson, 2000, Zhang and Thompson, 2004) for the English language. To measure passive vocabulary knowledge a yes/no test, based on these types of tests designed by Meara and his team (Meara and Buxton, 1987, Meara, 2010), was chosen. Besides, data relating the academic performance of the students in all the subjects of the curriculum and, especially in EFL, both in their third and fourth year of compulsory education were obtained from the schools.

## 5. Results

### *5.1. Reliability*

In order to be able to analyse the reliability of the sample obtained (RQ1), we first performed a half-split test to check whether the learners' score was higher when they provided answers for the first half of the test. Thus, we would measure whether the students' performance was similar from beginning to the end. We found that, in spite of being a test that has be completed in 15 minutes, the teenage students in our sample did not perform less satisfactorily in the last part of the test. The statistical measure used was the Spearman-Brown coefficient and the value obtained, 0.773, was statistically significant ($p<.000$).

We also performed a Cronbach's alpha to check the consistency of answers from the students. We wanted to check whether they obtained consistent results, ranging from 0 when they provided no answer or a non-scoring word to 4 when they supplied 4 different scoring words for each of the cue words in the test. Again the valued obtained, 0.688, is statistically significant (p<.000) but clearly on the low side of the scale.

### *5.2. Correlation with other language proficiency measures*

To check whether our learners' Lex30 scores correlated with other language proficiency measures (RQ2), we analysed their relationship with the results of the two other language tests administered to the students, i.e. the passive vocabulary test and the grammar test, as well as with their global marks in their EFL school subject. As can be seen in table 2 below, the results of this analysis show moderate correlation indices with the three measures considered.

**Table 2**: Correlations among proficiency measures

| | | LEX30 (T1) | Passive Vocab. | Grammar test | Marks in English |
|---|---|---|---|---|---|
| Lex30 (T1) | Pearson Correlation | 1 | ,509** | ,420** | ,548** |
| | Sig. (2-tailed) | | ,000 | ,003 | ,000 |
| | N | 48 | 48 | 48 | 44 |
| Passive Vocab. | Pearson Correlation | | | ,537** | ,528** |
| | Sig. (2-tailed) | | | ,000 | ,000 |
| | N | | | 48 | 44 |
| Grammar | Pearson Correlation | | | 1 | ,614** |
| | Sig. (2-tailed) | | | | ,000 |
| | N | | | 48 | 44 |
| Marks in English | Pearson Correlation | | | | 1 |
| | Sig. (2-tailed) | | | | |
| | N | | | | 44 |

**. Correlation is significant at the 0.01 level (2-tailed).

We also looked at this relationship by collapsing these three measures into a single global measure, which would result from their addition. The results of this further analysis were not very different, with only 30% of the variance being explained by the results in Lex30 (r=.550; n=44; p<.000; $R^2$=.302), which given the close relationship between vocabulary and proficiency (Schmitt, 2010) may be considered as a weak correlation.

A more detailed analysis of the results of the two tests that measure vocabulary proficiency, both productive (Lex30) and receptive (Yes/No test) shows a surprisingly weak connection between them (r=.509, n=48; p<.000; $R^2$=.205). In fact, the relationship between Lex30 and passive vocabulary in our sample is not strictly linear, as at approximately the 80% of passive vocabulary level Lex30 scores do not only not increase but even show a decreasing tendency (see the distribution in Figure 2 as compared to the general distribution when related to the overall proficiency measure in Figure 1)

**Figure 1:** Lex30 and overall proficiency measure

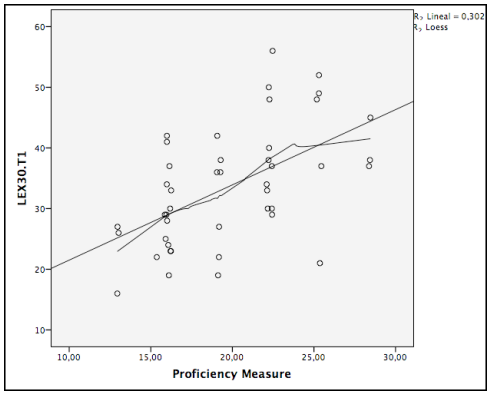
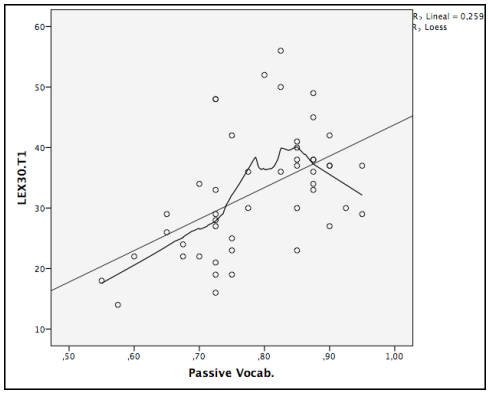



**Figure 2**: Lex30 and passive vocabulary




In contrast, Lex30 scores have a closer connection to another measure derived from the elicitation task used for the test: the number of words (i.e. tokens) the learners are able to produce when performing the associative task (Pearson's r= .885; n= 48; p<.000; $R^2$=0.783). This means that Lex30 scores are also closely linked to the number of non-scoring answers provided by the learners, which given that the elicitation task is timed (15 minutes) may be considered as an indication a connection with fluency.

### *5.3. Long-term vocabulary growth*

The results of the longitudinal study (the comparison of the learners' performance in the test with a time span difference of eighteen months, RQ3) confirmed the hypothesis that secondary school learners' productive vocabulary grew over a period of nearly two academic years (see Table 3):

**Table 3**: Comparison Lex30 results at T1 and T2

| | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| LEX30. T1 | 32,79 | 48 | 10,148 | 1,465 |
| LEX30.T2 | 37,42 | 48 | 10,671 | 1,540 |

Paired t-test: t=-3,413; df=47; p= 0.001

It is interesting to note, however, that the mean score at T2 only represents an increase of 14% with respect to T1.

These global results, however, disguise the situation of the individual schools studied. Thus, whereas school A does not show a statistically significant increase in Lex30 scores (35.37 at T1 vs. 36.26 at T2), school B scores are significantly higher at T2 (29,48 at T1 vs. 38.90 at T2 z=-3.599, p=000). In fact, given the different English levels of the two schools, we thought that this phenomenon could be related to the individual learners' language proficiency.

**Table 4**: Student T comparison of lower vs. higher level learners

| ENGLISH.LEVEL | | | Mean | N | Std. Dev. | Std. Error Mean | T | Sig (two-tailed) |
|---|---|---|---|---|---|---|---|---|
| Lower-level learners | Lex30 | T1 | 28,96 | 24 | 7,658 | 1,563 | -3,572 | ,002 |
| | | T2 | 34,71 | 24 | 10,145 | 2,071 | | |
| Higher-level learners | Lex30 | T1 | 38,24 | 17 | 8,800 | 2,134 | -1,558 | ,139 |
| | | T2 | 41,47 | 17 | 10,689 | 2,593 | | |

In order to confirm this possible explanation, we compared the vocabulary growth of lower-level and higher-level students, regardless of the school they were attending.

As shown in Table 3, lower level students in our sample demonstrated a statistically significant growth in their productive vocabulary whereas higher-level students did not.

### *5.4. Vocabulary group frequency*

Lex30 is not designed to make distinctions between the different vocabulary bands for two obvious reasons: 1) the scoring is based on the distinction between the non-scoring 1K band and the rest of frequency bands, which are awarded a point (cf. Fitzpatrick and Clenton, 2010: 548), and 2) the sample of words (a maximum of 120, but usually much smaller) does not easily permit to determine statistically significant differences. However, when the analysis is not restricted to the individual learner but, as in our case, it deals with the possibility of using the test to describe the vocabulary level of a group of secondary school learners, the sampling problems disappear.

The number of words at each frequency band is now large enough to allow for a comparison between the different frequency bands. Thus, even though it would not be possible to establish whether the frequency model works for individual learners, it is perfectly feasible to check whether it works for a group of learners.

To this end, instead of analysing individual frequency bands, we considered that it was more interesting to analyse the overall pattern by looking at vocabulary groups. As a consequence, following Schmitt and Schmitt (2012), we divided, the words produced by the students into three main groups: high frequency vocabulary (comprising 1k, 2k and 3k), mid-frequency vocabulary (from 4k to 8k), and low-frequency vocabulary (9k onwards). The results for both time 1 and time 2 are shown in table 5. As the results for the mid- and low-frequency groups were not normally distributed, we used the non-parametric Friedman's test to find whether the differences between the groups were significant within each testing time. The results of the test were performed separately for each of the times and were both statistically significant ($X^2$=73.661, asympt. Sig.=.000 for T2 and $X^2$=82.714, asympt. Sig.=.000 for T2) and the post-hoc comparisons found that all groups within the same testing time were different except for the mid vs. the low frequency groups, whose difference were not statistically significant at T1, although they were at T2.

**Table 5:** Vocabulary frequency groups at times 1 and 2

| | | N | Min. | Max. | Mean | Std. Deviation |
|---|---|---|---|---|---|---|
| T1 | High frequency | 48 | 23 | 104 | 55,44 | 19,792 |
| | Mid frequency | 48 | 2 | 15 | 6,52 | 3,255 |
| | Low frequency | 48 | 1 | 15 | 7,06 | 3,664 |
| T2 | High frequency | 48 | 38 | 96 | 67,25 | 16,703 |
| | Mid frequency | 48 | 4 | 19 | 9,65 | 4,066 |
| | Low frequency | 48 | 0 | 13 | 5,52 | 2,975 |

As expected, the results indicate a separation in the behaviour of the high frequency group and the other two groups of vocabulary.

## 6. Discussion

As stated above, the main aim of the study reported in this paper is to analyse the ability of Lex30 to determine the productive vocabulary of two groups of secondary school students focusing on four main aspects: first, the test reliability to measure the productive vocabulary of these learners; secondly, whether it correlates with general language proficiency; thirdly, whether it measures vocabulary growth over relatively long periods of time (eighteen months); and, fourthly, if it is sensitive to the possible effect of the context of learning on the productive vocabulary of the learner.

In relation to the first aspect, i.e. reliability (RQ1), this is, in our view, an important question when dealing with children's and teenagers' assessment. It has been pointed out (McKay, 2006) that their lack of cognitive maturity may affect their attention span, which may vary from one day to the other thus resulting in inconsistent results. Especially important in this sense is the short amount of time they are given to complete the test. Our data suggest that the students in our sample did not perform significantly less satisfactorily in the second half of the test despite the short amount of time that it needs to be completed. As a result, we may conclude that, even if the reliability scores are not very high, Lex30 seems to be a reliable instrument when used with young learners.

As for the importance of confirming the relationship between Lex30 scores and other language proficiency measures (RQ2), it should be noted that productive

vocabulary has consistently been demonstrated to be linked with language proficiency (Meara and Jones, 1988; Read, 2000; Schmitt et al, 2001; Zareva, Schwanenflugel and Nikolova, 2005; Crossley et al. 2011). Lex30 has not been an exception as it has been shown to distinguish between learners at different levels of proficiency (Fitzpatrick & Meara, 2004; Walters 2012). Walters (2012) found that Lex30 scores significantly discriminated among proficiency levels (high beginners, intermediate and advanced). In our data, we found statistically significant correlations between our learners' scores for Lex30 and the three other L2 proficiency measures, although they were moderate. Particularly unexpected was the weak connection between the two tests employed to measure the students' vocabulary knowledge (productive and receptive), especially, if we consider that Meara and Fitzpatrick (2000) found a correlation of 0.841 ($p<0.01$) between Lex30 and a yes/no test of passive vocabulary. A more in-depth analysis of the relationship of both tests results actually shows a non-strictly linear relationship, as shown in figure 2 above.

Simultaneous processes, not necessarily incompatible, may explain this lack of correlation between receptive and productive vocabularies. Thus, the demonstrated gap between productive and receptive vocabularies (cf. for example Meara, 2009), perhaps bigger in low-level learners, may provide one explanation, while the particular teaching context, CLIL, where the present study was carried out, may provide a complementary one. In parallel to what happens in immersion schools (e.g. Lyster, 1993; Swain, 1995; Swain and Lapkin, 1998), CLIL students may have developed more fully their receptive skills and the methodologies used in the CLIL classroom (teacher-fronted classes and little interaction in class) may have given a stronger impact on receptive vocabulary.

Interestingly enough, a much closer correlation was identified between the Lex30 scores and the number of words that the learners were able to produce when answering the test. These results may be an indication that, with low-level learners, the ability to provide a greater number of word types in Lex30, which could be associated to lexical diversity measures used in free production (cf. Milton, 2009), may be more representative of the productive vocabulary they have at their disposal. To a certain extent, when completing the test, learners are writing an 'associative text'. In this way, a lexical sophistication measure, i.e. the number of infrequent words, provided by the standard scoring of the test, would be a less appropriate measure to be able to make finer distinctions, especially if as the research has shown (cf. the review by Canga, 2013) Spanish secondary school learners possess, on average, a passive vocabulary that is not greater than the 1K band.

In relation to the third aspect analysed, i.e. the sensitivity of the test to determine a development in the learners' vocabulary size over long periods of time (RQ3), our data

shows that Lex30, indeed, measures vocabulary growth over relatively long periods of time (18 months). Our overall global scores show a 14% increase of the mean score at T2 with respect to T1. When the two schools are analysed separately, the picture obtained is, however, different with school A showing a non-statistically significant increase whereas school B scores are significantly higher at T2. Our analysis of higher and lower level student's performance, irrespective of their school, may be related to this finding. As shown, lower level students demonstrate a statistically significant growth in their productive vocabulary whereas higher-level students do not.

Two main explanatory lines may be established here. First, it could by hypothesized that the vocabulary of the high level learners studied has stopped growing and that there has been no progress in the number of words they are able to recall. In our opinion, this seems very unlikely as 1) the language environment to which students are exposed is very rich in vocabulary and 2) their general English knowledge, as their marks in this school subject indicate, point to the contrary. Besides, this could be considered as a possibility if their low-level classmates had made no progress, which was not the case.

A second explanation could be connected to the existence of certain ceiling effects in Lex30 scores. The high scores obtained at T1, especially by high-level learners, will make it difficult to make some progress when they are measured at T2. As stated in the methodology, using the JACET lists, the average score for the sample of our 48 students is 32.71. If we compare this score to the ones obtained in other studies on Lex30, particularly in school contexts (11.31 in Jiménez Catalán and Moreno Espinosa, 2005; and 19.15 in Moreno Espinosa, 2010) ours seems to be certainly high, even considering that the samples are by no means comparable since the learners in their sample were doing their fourth and sixth year of primary education respectively (9-10 and 11-12 years of age) whereas our learners at T2 were finishing their fourth year of secondary education (15-16 years of age). It should also be noted that the use of JACET 8000 lists in the scoring may not be wholly appropriate to score our sample as these list were designed having the Japanese school context in mind (cf. Uemura and Ishikawa, 2004).

Finally, the fourth aspect of our analysis was to check the possible sensitivity of the test to the context of learning. As with most vocabulary tests, the underlying construct of Lex30 is word frequency since it is hypothesized, following usage-based language acquisition theories, that the more frequent a word is the more likely it is to have been encountered in the input. This, in turn, involves that the most frequent words would be learned first, whereas lower frequency ones would appear later in the vocabulary of learners. As a consequence, one would expect that the number of words belonging to each frequency band would drop as we "move from higher to lower frequency levels" (Brown, 2012:21).

When the test used is Lex30, showing that this frequency model works at an individual level is not simple. Unlike other high-stakes vocabulary tests (e.g. PVLT), Lex30 is not designed to make distinctions between the different vocabulary bands. As pointed out above, the scoring is based on the distinction between the non-scoring 1K band and the rest of frequency bands, which are awarded a point and the sample of words does not easily permit the researcher to determine statistically significant differences. However, when the analysis is not applied to an individual learner but to a whole group the sampling problems disappear.

The results of our analysis showed a separation in the behaviour of the high frequency group and the other two groups of vocabulary which is in line with some of Aizawa's findings (2006, cited in Milton, 2009). According to this author, the frequency model seems to work well only with the first four frequency bands (1K-4K), after which the curve of the frequency model flattens out and "the frequency of words becomes more similar" (Milton, 2009:28).

However, it is interesting to note that the 'low frequency' group is the only vocabulary group in which we can find that the number of words produced by the students significantly decreases between T1 and T2 (Z=-3.210, asympt. Sig.=.002). In our view, this may be related to the particular learning context, a CLIL school, where the sample was gathered. It can be hypothesized that, given that certain CLIL courses focus on specialised terminology related to the specific subject matter that is being dealt with, the low frequency vocabulary group is over-represented in our sample, especially at T1.

In fact, if we examine some of the associations given by the students in our sample not found in responses native speakers gave to Lex30 cue words (cf. EAT), we can immediately see that there is a large group of specialised words, i.e. typical of non-language disciplines such as the Natural or Social Sciences (see Table 6 below).

**Table 6**: Specialised vocabulary in the samples

aids area assistant aware benefit channel communicate computer construction consume economy environment expert global globalisation goal ignorant injure injured injury instance investigate investigation location maintain mature medical military orientation periodic physical plus professional prohibition project psychological reaction requirement restriction role subordinate survival survive technology text thematic transport undertake vehicle virtual visual abundant acid adjective alcohol alien altitude aluminium anatomy anorak antennae anthropologist armchair artisan astrology astronomy atom bacteria bat biological biology blackberry bracket cable calcium candle carrot cartography

cartoon cast cauliflower cd cell cemetery ceramic cereal chef chew chocolate choke cholera cinema closet coconut column compliance consonant contest costume countryside crystal cucumber curriculum cytoplasm decoration demography dependence destiny diary dictatorship dioxide disaster disobedience documentary drill ecosystem electronic era excavation excrement excursion exocitosis expulse extortion faeces fantastic felicity fizzy flu fluid fossil fox fragile fridge gaseous genetic geographic grape graphic gym headache homicide horrible humorous hurricane hygienic idiom immunology impolite incisive inexperienced ingredient inorganic internet interview invasion jar kiwi landscape lava legend lemma lentil lesion lettuce lion lipid litter localization loo magazine magician mandarin mars math mathematic mathematics microbiology microscope mole mollusc monument morphology motorbike movie nitrogen nitrogenous notebook nutrient oak obligatory onion opera organic overcoat oxygen paella pasta pea pepper petrol phrase physiology pine pirate planet plastic pole pollution porcelain potato pox princess pronunciation protein pumpkin puppet racism raid recipient recollection remote repetitive retail robbery rugby sausage savannah scavenger shark shorts shotgun siege siesta smash strawberry subtract subtraction suitcase sulphur supermarket syllable tangerine territory terrorist textile theology topographic toxic toxicity tradesman trash tsunami tuberculosis tuna vase vegetal veracity virus vitamin volcano vomit wee witchcraft zoo

As a consequence, the use of a word association task where students can draw on any word that comes to their mind may result in an inflation of scores as we have repeatedly seen in the present article. This means that, although the frequency model has been shown to work well in instructional contexts and with certain vocabulary tests such as the Yes/No test (cf. Milton, 2009), it may be the case that, when the test target is precisely low frequency vocabulary and the instructional context, as happens with CLIL, focuses on specialised vocabulary, the frequency model may be disrupted and the results of the test less consistent because the learners' knowledge of technical words may result in inflated results.

## 7. Conclusion

The results of the study reported in the present article suggest that the design and implementation procedures of Lex30 could make it an appropriate test to be used with secondary school learners. However, they also seem to indicate that, especially in specific educational contexts such as a CLIL, Lex30 scores with such population should be interpreted with caution. The analysis of its reliability and validity with a population of CLIL secondary school students has revealed some elements that need to be taken into account when interpreting the results from this test: 1) although its

reliability is acceptable, it is not a highly reliable test; 2) the concurrent validity with other proficiency measures, especially with a passive vocabulary, is certainly low; 3) contrary to expectations and to the findings from other studies (cf. Fitzpatrick and Clenton, 2010), it is not very sensitive to showing vocabulary growth, especially when learners score high at the first point in time; and 4) The association task it uses to get a sample of the secondary school learners vocabulary may result, in certain contexts (those context in which they are widely exposed to technical vocabulary as in the case of CLIL), in an inflation of the number of words known by the students.

Further research, with a bigger sample, is needed to be able to confirm these preliminary conclusions, which would greatly benefit if a more appropriate benchmark corpus of native speaker secondary school learners could be used (cf. Fitzpatrick et al., 2015).

## Acknowledgements

The research reported in this paper was made possible by funding of the project *"Multilingüismo y multiculturalidad como factores positivos en el desarrollo cultural, afectivo y cognitivo del alumnado. Investigación sobre los efectos lingüísticos y psicosociales de las experiencias de inmersión (proyectos bilingües) y de no inmersión (alumnos inmigrantes) en el sistema educativo extremeño"* (reference: PRI08A127) by the Extremaduran regional government (*Junta de Extremadura*).

## References

Alderson, J.C. 2000. "Technology in testing: the present and the future". *System* 28 (4): 593–603.

Alejo González, R. and Piquer Píriz, A.. 2010. "CLIL teacher training in Extremadura: A needs analysis perspective". In D. Lasagabaster and Y. Ruiz de Zarobe (eds) 2010 *CLIL in Spain. Implementation, Results and Teacher Training.* Newcastle, UK: Cambridge Scholars Publishing.

Baba, K. 2002. "Test review: Lex30". *Language Testing Update* 32: 68-71.

Bachman, L. F .1990. *Fundamental considerations in Language Testing* Oxford: Oxford University Press.

Brown, D. 2012. "The frequency model of vocabulary learning and Japanese learners". *Vocabulary Learning and Instruction* 1(1): 20-28.

Canga, A. 2013. "Receptive vocabulary size of secondary Spanish EFL learners". *Revista de Lingüística y Lenguas Aplicadas* 8: 66-75.

Crossley, S. A., Salsbury, T. and McNamara, D. S. 2011. "Predicting the proficiency level of language learners using lexical indices". *Language Testing,* 28 (4): 561-580.

EAT. 1973. Kiss, G.R., Armstrong, C., Milroy, R., and Piper, J. "An associative thesaurus of English and its computer analysis". In A.J. Aitken, R.W. Bailey and N. Hamilton-Smith (eds) *The Computer and Literary Studies*. Edinburgh: University Press.

European Commission. 2012. *First European survey on language competences: Final report*. European Union.

Fitzpatrick, T. 2007. "Productive vocabulary tests and the search for concurrent validity". In H. Daller, J. Milton and J. Treffers-Daller (eds). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

Fitzpatrick, T. and Clenton, J. 2010 "The challenge of validation: assessing the performance of a test of productive vocabulary". *Language Testing* 27 (4): 537- 554.

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

Fitzpatrick, T., D. Playfoot, Wray, A., Wright, M. 2015. "Establishing the reliability of word association data for investigating individual and group differences". *Applied Linguistics* 36 (1): 23-50.

JACET. 2003. *JACET list of 8000 basic words*. Tokyo: Japan Association of College Teachers.

Jiménez Catalán, R. Mª. and Moreno Espinosa, S. 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.

Kane, M. (2012). "Validating score interpretations and uses". *Language Testing* 29 (1): 3-17.

Koizumi, R. 2004. "Validating a productive vocabulary knowledge test for novice Japanese learners of English". *JACET summer seminar proceedings (Facilitating Vocabulary Acquisition) 4*: 64-67.

Laufer, B. 1998. "The development of passive and active vocabulary in a second language". *Applied Linguistics* 19: 255-271.

Laufer, B. and Goldstein, Z. 2004. "Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness". *Language Learning* 54: 469-523.

Laufer, B. and Nation, P. 1995. "Vocabulary size and use: Lexical richness in L2 written production". *Applied Linguistics* 16 (3): 255-271.

_____. 1999. "A vocabulary size test of controlled productive ability". *Language Testing* 16: 33-51

Lyster, R. (1993). *The effect of functional-analytic teaching on aspects of sociolinguistic competence: A study in French immersion classrooms at the Grade 8 level*. University of Toronto: Toronto, Ontario, Canada

Lorenzo, F., Casal, S., and Moore, P. (2010). "The effects of content and language integrated learning in European education: key findings from the Andalusian sections evaluation project". *Applied Linguistics* 31: 418–42.

McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam/Philadelphia: John Benjamins.

Meara, P. 2010. [2nd edition]. *EFL Vocabulary Tests*. College Swansea: Centre for Applied Language Studies.

Meara, P. and Bell, H. 2001. "P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts". *Prospect* 16 (3): 323-337

Meara, P., and Buxton, B. 1987. "An alternative to multiple choice vocabulary testing". *Language Testing* 4 (2): 142-154.

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

Meara, P., and Fitzpatrick, T. 2004. *Lex30 v 2.0*. Swansea: Lognostics.

Meara, P., and Jones, G. 1988. "Vocabulary size as a placement indicator". In P. Grunwell (ed) 2008. *Applied linguistics in society*. London: Center for Information on Language Teaching and Research

Meara, P. M., and Olmos Alcoy, J. C. (2010). "Words as Species: An Alternative Approach to Estimating Productive Vocabulary Size". *Reading in a Foreign Language* 22 (1): 222-236.

Milton, J. 2009. *Measuring Second Language Vocabulary Acquisition*. Bristol/ Buffalo/ Toronto: Multilingual Matters.

Moreno Espinosa, S. 2009. "Young learners' L2 word association responses in two different learning contexts". In R. M. Jiménez Catalá and Y. Ruiz de Zarobe (eds) *Content and Language Integrated Learning in Europe* (pp. 241-256). Clevedon: Multilingual Matters.

____. 2010. "Boys' and girls' L2 word associations". In R. M. Jiménez Catalán (ed) *Gender Perspectives on Vocabulary in Foreign and Second Languages*. Basingstoke: Palgrave Macmillan. 139-163.Nation, P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House

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

Ruiz de Zarobe, Y. and Jiménez Catalán, R. (eds.). 2009. *Content and Language Integrated Learning. Evidence from Research in Europe*. Bristol: Multilingual Matters.

Schmitt, N. 2010. *Researching Vocabulary. A Vocabulary Research Manual.* Basingstoke, Hampshire: Palgrave.

Schmitt, N., and Schmitt, D. 2012. "A reassessment of frequency and vocabulary size in L2 vocabulary teaching". *Language Teaching*: 1-20.

Schmitt, N., Schmitt, D., and Clapham, C. 2001. "Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test". *Language testing*, 18 (1): 55-88.

Swain, M. 1995. Three functions of output in second language learning. In G. Cook and B. Seidlhofer (eds) *Principle and practice in applied linguistics: Studies in honour of HG Widdowson*. Oxford: Oxford University Press.

Swain, M., and Lapkin, S. 1998. "Interaction and second language learning: Two adolescent French immersion students working together". *Modern language Journal*, 82: 320-337.

Telchrow, J. M. 1982. "A survey of receptive versus productive vocabulary". *Interlanguage Studies Bulletin* 6: 3-33.

Uemura, T. and Ishikawa, S. 2004. "JACET 8000 and Asia TEFL vocabulary iniatiative". *Journal of Asia TEFL* 1: 333-347.Walters, J. 2012. "Aspects of Validity of a Test of Productive Vocabulary: Lex30". *Language Assessment Quarterly* 9: 172-185.

Wolter, B. 2002. "Assessing proficiency through word associations: is there still hope?" System 30: 315-329.

Zareva, A., Schwanenflugel, P. and Nikolova, Y. 2005. "Relationship between lexical competence and language proficiency". *SSLA* 27: 567–595.

Zhang, S. and Thompson, N. 2004. "DIALANG: A Diagnostic Language Assessment System (review)". *The Canadian Modern Language Review* 61(2): 290-293