Research Papers in Language Teaching and Learning
Vol. 11, No. 1, February 2021, 159-172
ISSN: 1792-1244
Available online at http://rpltl.eap.gr

Validating Pic-Lex for pre-primary and primary school age children

Chloe MILLS & James MILTON

Vocabulary assessments are widely used by researchers, teachers, and clinicians, and it is vital that any assessment used in practice is informed by research. This study trials Pic-Lex (Alexiou, 2019), a new picture-based receptive vocabulary test for children speaking English to establish whether the results it produces appear reliable and valid. The participants were 40 schoolchildren aged from 3 years and 2 months to 7 years and 9 months. Validation was assessed using an argument-based framework. Pic-Lex was trialled alongside the British Picture Vocabulary Test (BPVS) and a high correlation was found between Pic-Lex and raw BPVS scores (r = 0.79, p = <.01). The receptive vocabulary sizes of this sample averaged around 4000–5000 words, which is similar to other figures from the literature. The development of this new research-based receptive vocabulary test can have positive implications for vocabulary interventions in several contexts.

Key words: vocabulary, receptive size, children, EFL, test, assessment, Pic-lex

---

1. Background

Vocabulary is a key aspect of children's language development. Low levels of vocabulary can lead to educational struggles in school (Biemiller & Slonim, 2001; Milton & Treffers-Daller, 2013) while also impacting wellbeing and mental health (Oxford University Press, 2018). These factors, along with the discourse surrounding the idea of a vocabulary "gap" (Quigley, 2018), show that there is a pressing need to be able to accurately evaluate a child's vocabulary size. However, previously obtained vocabulary size estimates for children vary greatly due to the different methodologies employed by researchers and the relative lack of a reliable measure of vocabulary *size* for children.

The aim of this study is to trial a new receptive vocabulary size test on a small sample of British children. This new test, called Pic-Lex (Alexiou, 2019; for a review of the test see Alexiou & Milton, 2020), is designed to model the vocabulary acquisition process of young, pre-literate learners. In performing this preliminary piece of validation research and bearing in mind the recent call for greater rigour for the field of vocabulary assessment (Schmitt, Nation, & Kremmel, 2020), Pic-Lex will

be assessed in several areas of validity in order to discover (1) whether it is working as expected and (2) whether the results are meaningful.

1.1. Vocabulary size and the word gap

Vocabulary size and growth in children has been a topic of interest for researchers for decades. Despite this, vocabulary size estimates vary due to differing methodologies, from audio recordings (Hart & Risley, 1995; Hoff, 2003; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991) to vocabulary estimation through the use of bespoke vocabulary tests (Anglin, 1993; Biemiller, 2005; Biemiller & Slonim, 2001). On average, the literature suggests that children acquire anywhere between 800–3000 words per year. There is a rule of thumb that a rough vocabulary size can be calculated with the formula *(age – 2) x 1000* (Nation & Anthony, 2016). Conservative estimates of children's vocabulary size appear to confirm this, suggesting that eight-year-old children may know an average of 5000 root word meanings (Biemiller & Slonim, 2001). Based on an examination of school books, Anderson and Nagy (1993) suggest that children acquire 2000–3000 words per year. Anglin (1993) suggests that children in first grade (7–8 years old) know an average of 3000 root word meanings, but that this grows to 40,000 when you include inflected words, derived words, or idioms. Unfortunately, these studies all use different methodologies, ranging from unique vocabulary tests to extrapolation from educational materials—and these figures do not include the many studies examining vocabulary size via oral input. This variation means that even when promising results are obtained, they are not directly comparable to one other. For example, some of the variation in the vocabulary size figures might be explained by whether a researcher has chosen to present the test words in writing or orally, or both, and the degree to which the test words are contextualised.

While a difference of 1000–2000 words on either side of the average may be expected, concerns have been raised over the so-called vocabulary "gap" that may exist between children of different proficiencies in vocabulary. The "word gap" describes the difference in vocabulary between children who enter school with an expected vocabulary for their age and those who enter school with a vocabulary lower than expected for their age. This gap may be due to different socioeconomic backgrounds (Hart & Risley, 1995) and while some of this research may be methodologically flawed, a wealth of research supports the idea of individual differences affecting vocabulary size (Hoff, 2003). The word gap may affect progress in all school subjects and also impacts behaviour (Oxford University Press, 2018). In the UK, 21.2% of primary school pupils speak a language other than English in their home (Department for Education, 2019). The word gap may be particularly pronounced in these students who speak an additional or different language at home.

Vocabulary sizes need to be assessed if we are to investigate any potential word gap or if we want to track vocabulary development over time. The easiest way to do this is with a vocabulary size test. However, despite an increased focus on the importance of vocabulary, the proliferation of longitudinal vocabulary studies, and the increase in EAL pupils, there is still a lack of vocabulary measures that can be used with very young native English speakers (Nation & Coxhead, 2014). This is because designing a vocabulary test is complicated, especially for pre-literate children. It is vitally important to consider the methodology behind a vocabulary test yet many pieces of research do not seem to take this into account, resulting in the wide variety of vocabulary sizes reported in the literature. The issues that need to be addressed include how to define a word, the corpus and word lists used, how to test, what is the underlying construct, and test format and design (Nation, 2016; Schmitt, 2010). These choices have to be made in a principled, transparent manner and justifications should be given for the decisions made (Schmitt et al., 2020).

1.2. Current vocabulary size tests

The most common measure of receptive vocabulary size of children is the *Peabody Picture Vocabulary Test* (Dunn & Dunn, 2007). The PPVT can be used to examine whether a child struggles with receptive vocabulary (i.e., understanding words), rather than productive vocabulary (Groth-Marnat & Wright, 2009). The PPVT compares an individual child to standardised, normed figures. The norms are based on a large sample (over 6000 people), which is representative of the population of the United States. It is relatively easy and quick to use, but some level of training is required in order to administer and interpret the results effectively. Studies have shown the reliability and validity of results from the PPVT (Bracken & Murray, 1984; Stockman, 2000). The disadvantage of the PPVT is that the test items are based on *perceived* difficulty, rather than being linked to a model of children's language development. It also does not provide a vocabulary size estimate and only allows a researcher to compare an individual to generalised norms. Due to this, we cannot use the PPVT to model the learning process of vocabulary in children or investigate when particular words are learned. Furthermore, the PPVT has been criticised for a lack of ability to accurately judge the vocabulary knowledge of different demographic group, and it has been validated against standardised tests of intelligence and academic achievement but not in relation to aspects of semantic knowledge (Stockman, 2000). Therefore, the PPVT is unsuitable for gathering vocabulary size scores.

One test that may be the most suitable for younger children, with a principled and evidence-based design, is the Picture Vocabulary Size Test (Nation & Anthony, 2016). This is a multiple-choice test designed for young pre-literate speakers up to eight years old. In trials, it has shown itself to be suitable for children between six and eight years old; it presents a ceiling effect with children older than eight, yet is too difficult for those younger than six. Thus, there is still a lack of a test that can begin to the measure the vocabularies of pre-literate children younger than six; Pic-Lex intends to fill this gap.

1.3. Test validation

Test validation is a complicated area, and a call for more thorough test validation procedures in vocabulary research has recently been made (Schmitt et al., 2020). It is vitally important to clearly define the purpose of the test and specify the proposed interpretations and use of test scores, as well as ensure the validation process is evidence-based (Schmitt et al., 2020). Historically, validation has been dominated by views that multiple types of evidence should support score interpretation. Now, more recent frameworks have been adopted that attempt to provide explicit guidance on framing validity arguments (Chapelle, 2012; Chapelle, Enright, & Jamieson, 2010; Kane, 2013). Thus, this research will follow an argument-based framework, incorporating evidence from several areas of validity in order to support the validity argument. The framework begins with a specification and identification of the interpretations and use of the scores, followed by an evaluation of the overall plausibility of these interpretations. The argument will be framed using the following six areas of evidence, or inferences defined by Chapelle, Enright and Jamieson (2010): domain definition/description, evaluation, generalisation, explanation, extrapolation, and utilisation.

Construct validity is the evidential basis for score interpretation (Messick, 1995). In this framework, it is represented by the 'explanation' area of evidence. It is difficult to measure due to the amount of variables that may impacting this measure of validity. All information gathered on a test can contribute to understanding the scores, but a test becomes more valid if the score interpretations align with a theoretical rationale (Messick, 1995, p.743); in other terms, a test is only as valid as the conclusions that can be justifiably drawn from its results (Read, 2000). Pic-Lex is only valid therefore if the scores can be reliably extrapolated into sensible vocabulary sizes.

Therefore, as well as expecting a vocabulary test to accurately assess the construct being tested (i.e. receptive vocabulary size), a 'good' vocabulary test will also provide scores that correlate with other vocabulary size tests, show acceptable measures of internal reliability (i.e. Cronbach's alpha score), and demonstrate a visible frequency effect. If these criteria are met, then reliable conclusions can be drawn from the results.

1.4. Summary, aims, and objectives

The lack of good vocabulary size tests is due to the difficulty of designing and administering vocabulary assessments that can measure a child's vocabulary size. Most researchers, then, resort to using either a test such as the PPVT, which cannot provide detailed information on the *size* of a child's vocabulary, or they design their own particular test, which does not enable comparison with other research. This paper describes the validation of a computerised test designed to be suitable for very young children that gives a receptive vocabulary size measure. Using an argument-based validity framework, this study attempts to provide evidence for its use as an instrument that can be used to measure receptive vocabulary sizes of native speakers. Based on the overarching aim of providing validity evidence for this new picture-based vocabulary test, the specific aims of this research are as follows:

1) Does Pic-Lex provide us with a good Cronbach's alpha score, demonstrating reliability?
2) Does Pic-Lex produce correlations with another measure of receptive vocabulary, namely the British Picture Vocabulary Scale?
3) What does this test tell us about the receptive vocabulary sizes of a small sample of British children, and do these numbers align with previous research?

2. Methods

2.1. Participants

The participants in this study were pre-primary and primary school age children. A total of five-year groups were tested, with around half the children from each class undergoing testing for a total of 40 children. The sample size of 40 children sufficient for a pilot study and enables statistical testing to be carried out. The oldest child was 7 years and 9 months and the youngest was 3 years and 2 months. Table 1 shows the distribution of age and sex across the classes. Three children spoke a language other than English at home (Chinese, Bengali, and Arabic).

Class	***n***	**Male**	**Female**	**Age range (years; months)**
Rising Threes	7	4	3	3;2–3;9
Nursery	7	4	3	4;1–4;9
Reception	9	4	5	4;10–5;9
Year 1	8	5	3	5;11–6;8
Year 2	9	5	4	6;10–7;9
Total	**40**	**22**	**18**	**3;2–7;9**

Table 1. Participant characteristics.

2.2. Instruments

2.2.1. Pic-Lex

As already discussed, test development must be based on critical analysis and justifiable decisions. A detailed description of the development process for Pic-Lex can be found in Alexiou and Milton (2020). In summary, Pic-Lex is a computer-delivered test, which is intended for use as a measure of receptive vocabulary size. It measures knowledge of English vocabulary from the first five 1000-word frequency bands (Kilgarriff, 1995). From each band, 20 words are chosen to form a 100-item test. The test is comprised solely of nouns. Pic-Lex, like several other picture-based vocabulary tests, asks the respondent to choose a picture that matches a spoken word. Pic-Lex is easy to administer to children and is suitable for younger learners, even those who are pre-literate. Key to the development of Pic-Lex was ensuring that a vocabulary size could be calculated. A score of 100 suggests that the subject may know up to 5000 words in English, so the calculation required to generate a vocabulary size is to multiply the child's score by 50. Figure 1 presents a screenshot of Pic-Lex. It shows the four picture options and the audio file that reads out the test item (the test administrator may also read the word out loud). The test can be completed in 15 minutes.

Figure 1. Pic-Lex.

The cognitive processes behind word knowledge are complex and not yet fully understood. Therefore, the theoretical rationale behind a vocabulary assessment needs to be carefully considered. The mechanisms that guide task performance in Pic-Lex are related to what we know about how children learn words and what word knowledge comprises. To align a test with findings from the literature, we can break it down into its component processes (Messick, 1995). For Pic-Lex, these are:

- a respondent first hears and parses a spoken word;
- they have to consider and process four visual representations of words;
- then they have to try and link these to semantic knowledge in the brain;
- finally, they must try and correctly match their understanding of the test word to one of the picture options.

These steps should align with the underlying cognitive processes behind this type of vocabulary assessment: establishing a form–meaning link. The form–meaning link is a well-established aspect of vocabulary learning (Nation, 2013) and so Pic-Lex taps into a fundamental component of word knowledge. The form–meaning link correlates well with other aspects of word knowledge, at least in adults (Milton & Fitzpatrick, 2013); it is important to note that Pic-Lex does not intend to assess productive vocabulary or vocabulary depth (e.g. knowledge of word parts, synonyms, or collocations).

2.2.2. BPVS

The British Picture Vocabulary Scale (third edition) is based on the fourth edition of the Peabody Picture Vocabulary Test (Dunn and Dunn, 2007). It is a test of receptive vocabulary suitable for children from 3 years to 16 years 11 months, and is norm-referenced and measured individually. The BPVS is arranged so that test items become progressively more difficult as participants progress, and consists of 14 sets of 12 test items each. The BPVS produces a raw score and standardised scores can be calculated using a provided table. As stated by the BPVS manual, the standardised scores can be considered as more useful as they allow a person's attainment to be placed on a scale and compared to other tests that have the same mean (100) and standard deviation (15). The limitations of the PPVT, and thus the BPVS, have been discussed in Section 1.2.

2.3. Procedure

The children in this study were tested with both instruments with a gap of at least an hour in between to minimise interference and fatigue. Both the raw and standardised scores from the BPVS were recorded. Scores from Pic-Lex were recorded as raw scores out of 100 (e.g. a score of 95 means the child made 5 errors) and then converted into vocabulary size measures (multiplied by 50). Data were recorded in Excel imported into SPSS Version 22 for statistical analysis.

3. Results

The argument will be framed using the following six areas of evidence: domain description, evaluation, generalisation, explanation, extrapolation, and utilisation (Chapelle et al., 2010), as recommended by Schmitt et al. (2020). A brief description of each domain precedes the evidence for that area.

3.1. Domain Definition

This inference refers to the rationale for the test design, and requires a description of the domain from which the vocabulary items have been sampled and whether the test will be representative of what we want to find out. A description of Pic-Lex is provided in Section 2.2.1, and its purpose is clearly stated as a "bespoke testing tool that tests vocabulary knowledge in a simple, easy, fast and appropriate way for the specific age group of very young, pre-school learners" (Alexiou & Milton, 2020, p. 111). The domain from which the vocabulary items have been sampled is representative of

what the children will be expected to know. The scores from Pic-Lex will be used for a) measuring the vocabulary sizes of young learners, and b) identifying learners with lower-than-expected vocabularies, who may be experiencing the word gap and thus need intervention.

The test format is appropriate for what the study is aiming to investigate. The use of a multiple choice test can have some limitations, such as guesswork, and the use of pictures also has to be carefully considered. However, for the purposes of its use, Pic-Lex is an adequate test of passive vocabulary knowledge, and the benefits of time and ease of use outweigh its limitations. The multiple choice format allows test developers to simplify a task and cover a large sample of words (Read, 2000). Furthermore, the pictures are useful and necessary when testing pre-literate children.

3.2. Evaluation

This inference requires an analysis of the test scores and scoring procedure. Table 2 presents some descriptive statistics for Pic-Lex, showing an overall mean of 87.8 and a slightly large standard deviation. Figure 2 presents the mean Pic-Lex score by year group. The scores were not normally distributed for every class. A significant finding was that Pic-Lex had a readily apparent ceiling effect with the older children. All the participants in the Year 2, Year 1 and Reception groups were able to complete Pic-Lex. The older groups scored very highly; the mean score for all the participants in the Year 2 and Year 1 groups (*n* = 17) was 98. Only one child in the upper years scored below 95. There was more variability in the Reception group but the scores still remained high (mean of 96.5, *n* = 9). The mean dropped to 86 in the Nursery group (*n* = 7).

	n	**Minimum**	**Maximum**	**Mean**	**Standard deviation**
Score	40	10	100	87.8	23.1

Table 2. Descriptive statistics for Pic-Lex scores.

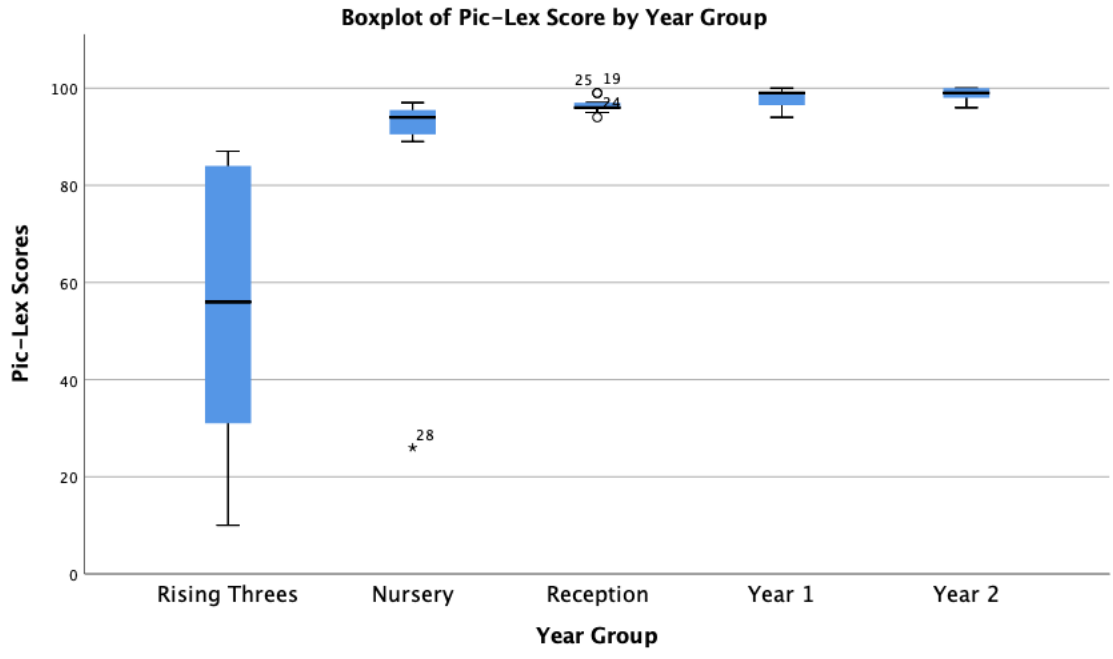

Figure 2. Boxplot showing Pic-Lex scores by year group.

The high scores achieved by the older children show a ceiling effect and demonstrate that the sensitivity of Pic-Lex is low for these age groups. Corroborating this is the fact that over half the test items were answered correctly across all levels. These results indicate that the items included in Pic-Lex may need adjustment if the test is to be administered to older children, as the ceiling effect will affect results and limits accurate evaluation of the children's lexicons.

The scores from the Rising Threes group present a different picture with much more variability. Four children in this group (n = 7) were unable to complete Pic-Lex. However, Pic-Lex is aimed at assessing the vocabulary of young, pre-literate children and so it must be accessible for this group. These results may suggest that Pic-Lex is unsuitable for this age group, as almost half the sample could not complete it; however, the sample is very small so no generalisation can be made. In contrast, the majority (6 out of 7) of participants in Rising Threes were able to complete the BPVS, which has a similar number of items depending on how far participants progress. This may raise the question: what are the differences between Pic-Lex and the BPVS and why can the youngest children complete the latter but not the former? Observation of the test participants suggested that as the BPVS increases in difficulty as participants progress, it holds the attention of the children for longer as it provides increasing challenge. On the other hand, in Pic-Lex the difficulty does not extend beyond the first 5000 words in English and so there is less challenge and the participants may lose interest. Other reasons could include the different interfaces of the test, or other design parameters such as the computerised nature of Pic-Lex versus the "offline" BPVS, or the pictures chosen. Further investigation is needed in this area and once again we can not draw conclusions based on this small sample.

The next aspect to consider is whether the results are related to frequency. The words in Pic-Lex are drawn from frequency bands (Kilgarriff, 1995), a decision that aligns with the literature guiding researchers to use an appropriate sampling procedure from different frequency bands when designing vocabulary tests (Read, 2000). Theoretically, people should know more words in the first 1000 words of English in comparison to later frequency bands as the first 1000 words are more common. The number of correct answers for each individual item in Pic-Lex was evaluated in order to determine the effect of frequency. The questions were combined into frequency bands, i.e. questions 1 to 20 are the first frequency band (1k) and so on. The average scores for each frequency band were then calculated. Table 3 shows the results for each frequency band as well as their mean rank.

Frequency band	Average % correct answers	Mean rank
1k	92.588	2.85
2k	94.999	2.87
3k	95.1	2.73
4k	95.141	3.13
5k	96.165	3.42

Table 3. Frequency data

The frequency profile was unexpected, as we would expect to see a decrease in average % correct answers as the frequency bands increase but instead we see an increase. The reasons for this may be that Pic-Lex only includes simple nouns. It is known that English-speaking children learn and use nouns before verbs (Saxton, 2017) and so a test that only includes nouns should assess a large proportion of the child's lexicon, but the restriction also means that Pic-Lex will not assess the entirety of a child's vocabulary knowledge (although this is difficult anyway). The inclusion of more parts of speech such as verbs or adverbs may improve the item and sampling validity and may potentially change the frequency results. But there was a clear rationale behind choosing only nouns. The small sample size may also have impacted the frequency results. The test needs further examination and adjustment in order to discover why this experiment did not present the intended frequency profile.

3.3. Generalisation

This section deals with the reliability and generalisability of the test scores. Cronbach's alpha is typically used to determine internal consistency; however, it may be ineffective in a vocabulary assessment situation, where each item may be considered its own separate construct (Schmitt et al., 2020). Despite this limitation, Cronbach's alpha is still a widely used concept of internal reliability. The Cronbach's alpha score of 0.789 demonstrates acceptable reliability. However, the hypothesis that participants would score higher on more common words (first frequency band) compared to less common words (later frequency bands) was not reflected in the results (see Table 3), which means that the internal consistency of Pic-Lex needs to be further investigated and a larger sample is needed to do this.

A one-way ANOVA was run (due to the nonparametric nature of the data) in order to determine the effect of age on Pic-Lex scores and see whether the test can differentiate between levels. This revealed a statistically significant difference between groups (F (4,35) = 8.61, p = < 0.05). A Tukey post-hoc test showed that the difference in means between the Rising Threes year group and the others was statistically significant (54.71 +/- 32.67, p = < 0.05), but there was no statistically significant difference between the mean scores of the other groups (p = .473).

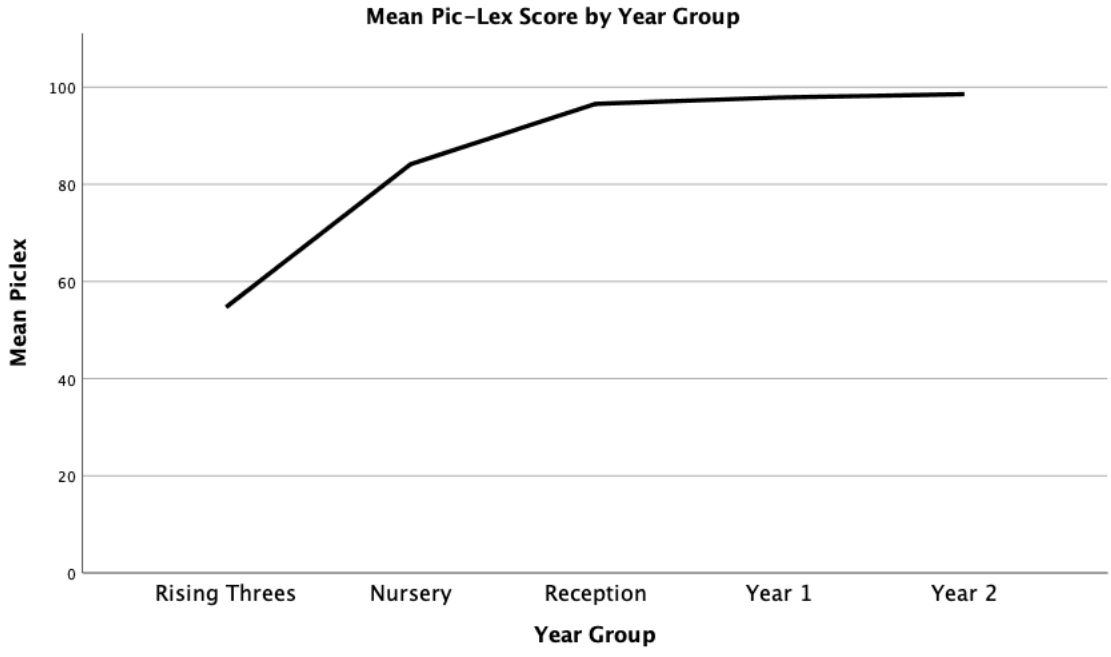

Figure 3. Means plot of Pic-Lex score by year group.

The best establishment of reliability would be a test–retest procedure. However, this was not possible within the constraints of this study. Thus, an important piece of future validation will be a further assessment of whether scores remain the same in a test-retest scenario, and further investigation into whether test-takers' scores can reliably distinguish between different groups.

3.4. Explanation

This inference involves linking the items and scores to the construct being tested which, in this study, is vocabulary knowledge. Here, we can attempt to quantify whether or not the scores can be converted into reasonable estimates of vocabulary size. We can also consider whether these sizes align with previous literature.

Table 4 and Figure 3 present the average vocabulary sizes for all the children who took part in Pic-Lex, including those who could not finish the test (*n* = 40). The error bars demonstrate the

Class	Age range (years; months)	*n*	Mean vocabulary size score
Rising Threes	3;2–3;9	7	2736
Nursery	4;1–4;9	7	4207
Reception	4;10–5;9	9	4828
Year 1	5;11–6;8	8	4894
Year 2	6;10–7;9	9	4928

Table 4. Mean vocabulary sizes, all children.

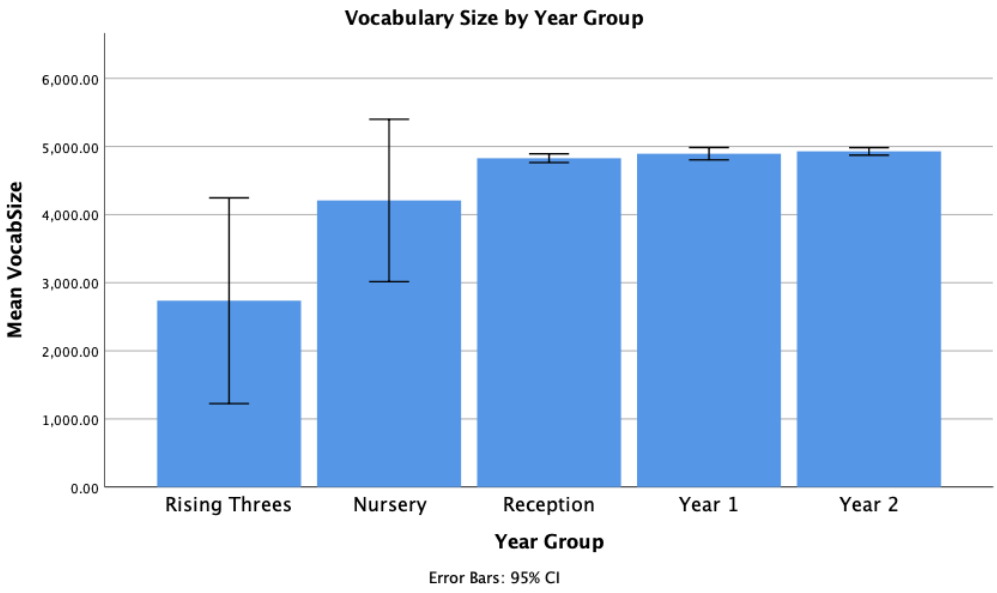

Figure 3. Mean vocabulary sizes including all children.

This figure also shows a ceiling effect, shown by the oldest year groups achieving close to the maximum score of 5000. As previously mentioned, the vocabulary of these groups is likely underestimated and not an accurate representation of their lexical knowledge. However, the numbers do broadly align with data from two previous studies (Anglin, 1993; Biemiller & Slonim, 2001) presented in Section 1.2. As this ceiling effect negates the estimation of vocabulary size in older children, if Pic-Lex is to be used with older age groups, the test developers could extend the frequency bands chosen and include more low-frequency words, in effect making the test "harder" and reducing the ceiling effect. However, this may not be necessary if Pic-Lex is only to be used with younger children.

The explanation area of validity also includes correlation of the test to another that tests the same construct. In Table 5, correlations between BPVS and Pic-Lex are displayed for all children (n = 40). The Pic-Lex scores and the standardised BPVS scores showed a moderate correlation (r = 0.56, p = <.01 two-tailed) and the Pic-Lex scores and the raw BPVS scores showed a high correlation (r = 0.79, p = <.01, two-tailed).

Paired Samples		**Correlation**	**Sig.**
BPVS Raw & Pic-Lex	0	.785**	.000
BPVS Standardised & Pic-Lex	0	.564**	.000

Table 5. Correlations between BPVS and Pic-Lex.

Concurrent validity determines the extent to which scores from different test instruments relate to each other. Pic-Lex shows a correlation with the British Picture Vocabulary Scale and in particular, the raw BPVS scores show a higher correlation to Pic-Lex scores than the standardised BPVS scores. This is most likely due to the fact that Pic-Lex is not a standardised test. Future research comparing Pic-Lex to other validated vocabulary tests would gather further evidence for concurrent validity.

Factors that may influence the score interpretations can also be considered here, such as test-taking behaviour. A large amount of guesswork was not observed in this trial. However, as in any multiple choice test, guessing on some questions is likely and may lead to influencing scores. However, research has shown that blind guessing is actually used as a last resort and that most participants know the meaning of the words, or have partial knowledge and use strategies that indicate this (Gyllstad, Vilkaitė, & Schmitt, 2015).

Finally, we can note that the pictures in Pic-Lex were recycled, i.e. used more than once, and so this may have influenced results due to children remembering that a picture has already matched with a previous word. This could be remedied very simply by gathering more pictures and ensuring they are only used once throughout the test.

3.5. Extrapolation

This step links test scores to a candidate's knowledge *outside* of the test-taking situation. This step does not have any evidence, as Pic-Lex was only run alongside the BPVS. While the above step provided evidence for knowledge of the same construct, in order to provide more a more complete

picture surrounding the candidate's ability outside of the test situation, future validation studies may run Pic-Lex alongside, for example, reading comprehension tests, in order to determine whether Pic-Lex is assessing the correct construct (Schmitt et al., 2020).

3.6. Utility and Impact

Finally, the utility and impact of the test scores must be discussed. As this is only a preliminary study, the full impact cannot be addressed, but the potential utilisation and impact can be considered. For example, if Pic-Lex is further improved and rolled out, and documentation is provided on the meaning of scores based on evidence gathered over several validation studies (e.g. a score < 60 means a child may need interventions), then Pic-Lex has the potential to be used for its intended use, i.e. in classroom settings in order to identify children who may need additional help in their English.

4. Conclusions

If we are to achieve the rigour that is needed in the field of vocabulary assessment, it is important to transparently present the facts regarding new vocabulary tests. In carrying out this research, it has become clear that so far the validation of Pic-Lex in its current form, presents a mixed argument. Pic-Lex has several strengths, including its ease-of-use, its format, its principled word selection, its reliability (as show by the Cronbach's alpha score), and its correlation with the BPVS. It provides most validation evidence in the area of explanation, and there is some evidence for validity in the areas of domain definition, evaluation and generalisation. This evidence allows us to be reasonably confident in the interpretation of generated scores.

Therefore, we can say that Pic-Lex was able to assess the receptive vocabulary sizes of young learners, working particularly well for 4- and 5-year-olds. The first two aims (a good Cronbach's alpha score, and correlations with another test of receptive vocabulary) have both been met. For the latter aim, we can provide general vocabulary size measures of around 4800 words for children aged 5–7; a higher sample size and some evaluation of test items is needed if more specific and accurate numbers are to be gathered. Finally, detailed SES data were not able to be collected in this research. We can only make some general observations, in that the school was in an affluent area and there were few children receiving free school meals (a marker of low-SES in the United Kingdom). Therefore, the vocabulary sizes collected can be said to be representative of a high-SES area of the United Kingdom.

However, the current Pic-Lex test has some weaknesses. The limitations lie mainly in its item selection, which led to a ceiling effect (if administered to older learners) and an effect on frequency scores. These weaknesses will need to be addressed in future research (see for more on this the updated version of Pic-lex in Alexiou, this volume). Also, because validation is an ongoing process (Fulcher & Davidson, 2007), it will not be possible to demonstrate a final validity argument, therefore, future studies will aim to provide ongoing validation evidence for Pic-Lex. Upcoming pilots and trials in different contexts (e.g. multilingual environments, EAL learners, low-SES groups) will identify further refinements to Pic-Lex before its eventual release into the public domain.

References

Alexiou, T. (2019) *Pic-lex: Picture Vocabulary size test.* http://gp.enl.auth.gr/piclex/

Alexiou, T. & Milton, J. (2020) Pic-Lex: A new tool of measuring receptive vocabulary for very young learners. In Zoghbor, W. & Alexiou, T. *Advancing ELT Education* (pp.103-113). Dubai, UAE: Zayed University publications, https://www.zu.ac.ae/main/en/research/publications/_books_reports/2020/AELE_Book_ALLT_ZU_Web_V02.pdf

Anderson, R. C., & Nagy, W. E. (1993). The vocabulary conundrum. *Center for the Study of Reading Technical Report, no. 570*.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, i-186.

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In A. Hiebert & M. Kamil, (Eds.). *Teaching and learning vocabulary: Bringing research to practice* (pp. 223-242). Mahwah, NJ: Erlbaum.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of educational psychology, 93*(3), 498.

Bracken, B. A., & Murray, A. M. (1984). Stability and predictive validity of the PPVT-R over an eleven month interval. *Educational & Psychological Research*.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…. *Language testing, 29*(1), 19-27.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational measurement: Issues and practice, 29*(1), 3-13.

Department for Education. (2019). *Schools, pupils and their characteristics: January 2019*. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/812539/Schools_Pupils_and_their_Characteristics_2019_Main_Text.pdf

Dunn, L., & Dunn, D. (2007). *Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4)*. Minneapolis, MN: NCS Pearson Psychcorp.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*: Routledge New York.

Groth-Marnat, G., & Wright, A. (2009). *Handbook of personality assessment* . Hoboken, NJ: Wiley.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics, 166*(2), 278-306.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*: Paul H Brookes Publishing.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development, 74*(5), 1368-1378.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: relation to language input and gender. *Developmental psychology, 27*(2), 236.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448-457.

Kilgarriff, A. (1995). BNC database and word frequency lists. Retrieved from http://www.kilgarriff.co.uk/bnc-readme.html

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist, 50*(9), 741.

Milton, J., & Fitzpatrick, T. (2013). *Dimensions of vocabulary knowledge*: Macmillan International Higher Education.

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review, 4*(1), 151-172.

Nation, P. (2013). *Teaching & learning vocabulary*: Boston: Heinle Cengage Learning.

Nation, P. (2016). *Making and using word lists for language learning and testing*: John Benjamins Publishing Company.

Nation, P., & Anthony, L. (2016). Measuring vocabulary size. *Handbook of Research in Second Language Teaching and Learning, 3*, 355-368.
Nation, P., & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching, 47*(3), 398-403.
Oxford University Press. (2018). *Why Closing The Word Gap Matters: Oxford Language Report*.
Quigley, A. (2018). *Closing the vocabulary gap*. Routledge.
Read, J. (2000). *Assessing Vocabulary*: Cambridge University Press.
Saxton, M. (2017). *Child language: Acquisition and development*. Sage.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.
Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching, 53*(1), 109-120.
Stockman, I. J. (2000). The new Peabody Picture Vocabulary Test—III: an illusion of unbiased assessment? *Language, speech, and hearing services in schools, 31*(4), 340-353.

Chloe Mills (822645@swansea.ac.uk) is a PhD student of Applied Linguistics at Swansea University studying first language vocabulary development in school children.

James Milton (email2@email.com) is Professor of Applied Linguistics at Swansea University, UK. A long-term interest in measuring lexical breadth, and establishing normative data for learning and progress, has led to extensive publications including *Measuring Second Language Vocabulary Acquisition* (Multilingual Matters, 2009).
