Routledge
Taylor & Francis Group

International Journal of Bilingual Education and Bilingualism

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rbeb20

Using the peabody picture vocabulary test in L2 children and adolescents: effects of L1

Claire Goriot, Roeland van Hout, Mirjam Broersma, Vanessa Lobo, James M. McQueen & Sharon Unsworth

To cite this article: Claire Goriot, Roeland van Hout, Mirjam Broersma, Vanessa Lobo, James M. McQueen & Sharon Unsworth (2021) Using the peabody picture vocabulary test in L2 children and adolescents: effects of L1, International Journal of Bilingual Education and Bilingualism, 24:4, 546-568, DOI: 10.1080/13670050.2018.1494131

To link to this article: https://doi.org/10.1080/13670050.2018.1494131

Published online: 19 Aug 2018.

INTERNATIONAL JOURNAL OF BILINGUAL EDUCATION AND BILINGUALISM
2021, VOL. 24, NO. 4, 546–568
https://doi.org/10.1080/13670050.2018.1494131

Routledge
Taylor & Francis Group

Using the peabody picture vocabulary test in L2 children and adolescents: effects of L1

Claire Goriot [a], Roeland van Hout[a], Mirjam Broersma[a], Vanessa Lobo[a,b], James M. McQueen[c,d] and Sharon Unsworth[a]

[a]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands; [b]Varendonck College, Asten, the Netherlands; [c]Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University, Nijmegen, the Netherlands; [d]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

ABSTRACT

This study investigated to what extent the Peabody Picture Vocabulary Test (PPVT-4) is a reliable tool for measuring vocabulary knowledge of English as a second language (L2), and to what extent L1 characteristics affect test outcomes. The PPVT-4 was administered to Dutch pupils in six different age groups (4-15 years old) who were or were not following an English educational programme at school. Our first finding was that the PPVT-4 was not a reliable measure for pupils who were correct on maximally 24 items, but it was reliable for pupils who performed better. Second, both primary-school and secondary-school pupils performed better on items for which the phonological similarity between the English word and its Dutch translation was higher. Third, young unexperienced L2 learners' scores were predicted by Dutch lexical frequency, while older more experienced pupils' scores were predicted by English frequency. These findings indicate that the PPVT may be inappropriate for use with L2 learners with limited L2 proficiency. Furthermore, comparisons of PPVT scores across learners with different L1s are confounded by effects of L1 frequency and L1-L2 similarity. The PPVT-4 is however a suitable measure to compare more proficient L2 learners who have the same L1.

ARTICLE HISTORY
Received 22 March 2018
Accepted 25 June 2018

KEYWORDS
PPVT; second language acquisition; cognates; lexical frequency; receptive vocabulary; assessment

Vocabulary tests are frequently used in research on monolinguals, and early bilinguals and second language (L2) learners, both to measure children's vocabulary development in a specific language, and to evaluate their overall language abilities (e.g. Bialystok et al. 2010; Dongsun, Yoon, and Jiyeon 2016; Poarch and van Hell 2012b). As previously noted (Gathercole, Thomas, and Hughes 2008), most often these tests have been developed for use with native speakers of the language being assessed, and therefore have been normed on a monolingual population. The way in which learners acquire an L2, however, by definition differs from how they acquire a first language (L1), in terms of for example context or age. The use of an L1 vocabulary test with an L2 population may therefore be problematic: Various L1 and L2 factors may have an influence on test outcomes, such as linguistic overlap between the L1 and L2 or how frequently young L2 learners encounter certain words. The question we address in this paper is to what extent an L1 vocabulary test can be reliably used with young L2 learners.

For children acquiring English as an L1, many language proficiency tests are available (for a discussion see Gathercole, Thomas, and Hughes 2008), one of the most commonly-used being the

CONTACT Claire Goriot c.goriot@let.ru.nl
This article has been republished with minor changes. These changes do not impact the academic content of the article.

Peabody Picture Vocabulary Test (PPVT), a receptive vocabulary test (Dunn 1959). This test has been widely used with monolingual English speakers, and with bilingual children in non-instructed settings (see for example Bialystok et al. 2010 for a meta-analysis). The use of the PPVT (or its British-English equivalent the BPVS) in L2 acquisition contexts is widespread, too (see for example Cohen 2016; Crevecoeur, Coyne, and McCoach 2014; Dahl and Vulchanova 2014; Dongsun, Yoon, and Jiyeon 2016; Jensen 2017; Leśnieuwska and Pichette 2016).

One of the domains in which the PPVT is frequently used with L2 learners is the domain of early foreign-language education. Also, in various European countries, using English as the language of instruction in addition to the official language in education has gained in popularity in recent years, both in primary (Huang 2016) and secondary education (Nikula 2017). Many researchers have investigated whether these educational programmes lead to gains in English receptive vocabulary. Such studies have been conducted in many countries, including Belgium (Buyl and Housen 2014), Finland (Merisuo-Storm 2007), France (Cohen 2016), Germany (Steinlen and Piske 2013), the Netherlands (Admiraal, Westhoff, and de Bot 2006; Lobo 2013; Unsworth et al. 2015; van der Leij, Bekebrede, and Kotterink 2010), Norway (Dahl and Vulchanova 2014), and Spain (Jimenez Catalan and Terrazas Gallego 2005). Many of them made use of the PPVT or the BPVS, either in its original form (Buyl and Housen 2014; Cohen 2016; Dahl and Vulchanova 2014; Unsworth et al. 2015; van der Leij, Bekebrede, and Kotterink 2010) or back-translated from the Dutch version of the test to English (Lobo 2013). The PPVT was not designed as a task for L2 learners, however, and the pupils in those studies learned English in a context differing from the one in which L1 learners learned it, at an age at which their L1 had already developed substantially. Different factors, such as the frequency with which certain words are used in the L2 or the linguistic overlap between the L1 and L2, may play a role in their L2 vocabulary development and influence their scores on the PPVT. In this study, we investigate the extent to which, given these factors, the PPVT is suitable for use with young Dutch pupils learning English as an L2 via an educational programme.

The PPVT

The PPVT was originally developed as a measure of verbal intelligence (Dunn 1959). The English version of the PPVT is currently in its fourth edition. This edition consists of 228 items grouped in 19 sets of 12 items each, arranged in order of decreasing frequency, with increasing difficulty being assumed. An item consists of four full-colour pictures. The participant has to select the picture that best matches the orally presented word. The items include verbs, adjectives, and nouns. The words belong to one of 20 different content categories, like animals, actions, or emotions. According to the guidelines in the manual, the start set is dependent on the age of the participant. If the participant gives two or more incorrect responses in one set, an easier set is presented. The basal set is determined as the set in which maximally one incorrect response is given. After selection of the basal set, the main phase of testing begins. Testing ends when the participant makes more than seven incorrect responses within a single set, thereby reaching the ceiling set. The PPVT-4 was normed on a sample of 3540 people, representative for the population of the United States. Both split-half and test-retest reliability were consistently high, with coefficients higher than .90 for all age groups (Dunn and Dunn 2007).

It is not surprising that the use of the PPVT is so widespread: it requires no literacy skills or oral response, there is minimal risk for stress or perceived failure, it is appropriate for use with participants aged two and older, and the coloured pictures are perceived as appealing to children. The authors of the test consider it as a useful measure for the assessment of 'the extent and nature of a person's knowledge of standard American English words' (Dunn and Dunn 2007, 3), even for individuals whose L1 is not English (Dunn and Dunn 2007). However, as noted above, the PPVT was not developed for L2 learners. Its use for such a population may therefore not be completely unproblematic. A potential problem involves the possible interplay between the L1 and the L2, and this may influence outcomes (Wood and Pena 2015).

L1 effects when investigating L2 vocabulary

A common view in the L2 acquisition literature is that L2 learners do not begin from scratch when they start building up a lexicon in the new language. Instead, they rely on the knowledge and the concepts that they have in their L1 (Dijkstra and Van Heuven 2002; Kroll and Stewart 1994). One aspect of children's L1 that may affect their vocabulary acquisition and, in particular, their performance on the PPVT is L1 word frequency. The structure of the PPVT-4 is based on the idea that some words are more frequent than others, and that children will acquire more frequent words before less frequent words and hence are more familiar with more frequent words (Dunn and Dunn 2007). However, word familiarity may be different for L2 learners than for native speakers. Wood and Pena (2015) showed that the difficulty level of the items in the PPVT-4, as determined by their order in the test, was positively related to children's error scores, but this relation was stronger in English monolingual children than in Spanish L2 learners of English. Differences in word familiarity between L1 and L2 learners may be especially more likely for children who have limited L2 exposure, as is the case for Dutch children who are exposed to English either at school or via media (Lindgren and Muñoz 2013). Words that are frequent in English as an L1 are not necessarily frequent in English as an L2, but frequency measures for English as an L2 are not available. It is therefore difficult to determine to which specific words children are exposed, and with what frequency. While we thus have no suitable measure of the frequency with which words may have been experienced in *English* in the sample of Dutch children we test, we investigate whether performance on the PPVT depends on *Dutch* frequency. It is however likely that children who start acquiring a new language at school at a time that they already have acquired their first language will do so by making use of their L1 lexicon. Consequently, the frequency of the words in the L1 may be a more important predictor of word knowledge in the L2 (English) than the L2 frequency itself.

Children may also rely on their knowledge of the L1 by recognising the similarities between words in the L1 and the L2. This might be particularly helpful for cognates, which in this study are defined as words that show semantic overlap between two languages, as well as large similarities in spelling and/or sounds. Translation pairs that show orthographic and phonetic similarities but that have different meanings (i.e. 'false friends') are not included in this study. It is known that meaning and form overlap helps children derive the meaning of a word (Pérez, Peña, and Bedore 2010; Potapova, Blumenfeld, and Pruitt-Lord 2016), even when children have limited exposure to the L2 (Bosma et al. 2016). Children have also been found to process cognates faster than non-cognates (Brenders, van Hell, and Dijkstra 2011; Poarch and van Hell 2012a). Just like adults (Dijkstra et al. 2010), children seem to show a gradual cognate facilitation effect: they are more likely to know the meaning of an identical cognate item than of a non-identical cognate, although they are also able to derive the meaning of non-identical cognates (Bosma et al. 2016). Older children are better at recognising cognates than younger children, especially if the items are non-identical cognates (Bosma et al. 2016). Since the number of cognates is greatest in closely related language pairs (Schepens et al. 2013), and Dutch and English are both Germanic languages that are known to share a large proportion of cognates (Broersma 2009; Schepens et al. 2013), it is likely that the PPVT-4 will contain relatively many English words of which the phonological similarity is close to their Dutch translation equivalent. When administering the PPVT-4 to native Dutch children it is thus to be expected that they may well benefit from these items.

Indeed, it has been noted that the PPVT-4 contains cognates for Dutch-speaking children (Lobo 2013; Unsworth et al. 2015). The researchers mentioned that this may have helped monolingual children with word association and recognition. Previous research has also shown that the PPVT-3 and PPVT-4 contain Spanish-English cognates (Potapova, Blumenfeld, and Pruitt-Lord 2016; Wood and Pena 2015), and that Spanish-English bilingual children perform better on cognates than on non-cognates (Potapova, Blumenfeld, and Pruitt-Lord 2016). Furthermore, a recent study with adult L2 learners showed that the PPVT-4 contains more French-English than Polish-English cognates, and hence French L1 speakers obtained higher scores on the PPVT-4 than Polish L1 speakers (Leśniewska,

Pichette, and Béland 2018). Similarly, in a large-scale experiment which compared foreign-language learners in seven European countries (Lindgren and Muñoz 2013) children living in a country in which the language of schooling was linguistically close to English (i.e. Swedish and Dutch) performed better on English listening and reading tasks than children living in a country where the language of schooling was less close to English (i.e. Croatian and Polish). Whilst that experiment did not use the PPVT, it does suggest that children might rely on cognate knowledge when being tested in their L2. All these studies thus suggests that when investigating L2 English vocabulary using the PPVT children's scores might be influenced by cognates.

The current study

Any measure of children's L2 vocabulary may thus be influenced by the fact that the frequency of the L2 words does not have to be the same as the frequency of the words in that language as an L1, and by the degree to which translation equivalents of the items in a test overlap in phonological form. Research on these issues is however limited. The current study addresses these issues by investigating the extent to which word frequency and phonological overlap between item-translation pairs predict the performance on the PPVT-4 by Dutch children who have limited exposure to English. The goal of this study more generally was to investigate how suitable the PPVT-4 is for measuring vocabulary knowledge of Dutch children and adolescents at different stages of learning English as an L2.

We conducted three experiments. In Experiment 1, in order to investigate lexical frequency, we translated all items of the PPVT-4 from English to Dutch, and examined if Dutch and English lexical frequencies are correlated. We also investigated cognate status. Contrary to previous studies (Bosma et al. 2016; Pérez, Peña, and Bedore 2010; Potapova, Blumenfeld, and Pruitt-Lord 2016; Wood and Pena 2015), we used a continuous measure instead of an arbitrary cut-off point to determine phonological similarity between pairs of English items and their translations. We expected the frequencies to be closely related to each other, since previous research has shown that even unrelated languages show considerable overlap in frequency (Moscoso del Prado Martín et al. 2004). Since English and Dutch are linguistically close, we expected the similarity of the item-translation pairs to be high.

In Experiments 2 and 3, we investigated whether word frequencies and phonological similarity measures collected in Experiment 1 predicted respectively primary-school and secondary-school pupils' scores on the PPVT-4. The children were learning English via an educational programme and/or were exposed to English via media. In both experiments, we investigated, as in previous studies (Buyl and Housen 2014; Cohen 2016; Dahl and Vulchanova 2014; Lobo 2013; Unsworth et al. 2015; van der Leij, Bekebrede, and Kotterink 2010), whether pupils who attended an English programme at their school and those who did not differed in their performance on the PPVT-4.

Experiments 2 and 3 were used to test three hypotheses. The first hypothesis was that the reliability of the PPVT-4 would increase when administering it to pupils who had more experience with English. Across the board, the older pupils are expected to have more experience with English than the younger pupils, and therefore to reach higher sets of the PPVT-4. For the youngest pupils, on the other hand, testing may regularly stop after the first few sets, resulting in a floor effect. Therefore, the test may not reliably differentiate between young pupils' vocabulary abilities.

Our second hypothesis was that children would perform better on English items that are more similar in form to their Dutch translations. Furthermore, since previous research has shown that children are better at recognising similarities between words as they become older (Bosma et al. 2016), we expected that the relation between form similarity and performance would get stronger in older pupils.

Third, we hypothesised that there would be a decreasing L1 frequency effect and an increasing L2 frequency effect as pupils get older. We expected that for the younger pupils in particular, L1 (Dutch) frequency would affect their performance, because of their limited experience with English. Pupils in

Dutch early-English primary schools generally receive English lessons for maximally one hour per week (Jenniskens et al. 2017) and mainly in an educational setting, and hence there is more room for Dutch than for English frequency to play a role. We expected that for older children, English frequency would become a more important predictor of vocabulary performance, since older pupils have more experience with and exposure to English. They may be exposed to English more, both outside school (Lindgren and Muñoz 2013), and at school: in mainstream education, secondary-school pupils receive between two and four hours of English lessons per week, and in bilingual education 50% of the lessons is in English (approximately 10 h per week for pupils following pre-university training) (EP-Nuffic, n.d.).

Experiment 1: Lexical frequencies and cognate status

The aim of this experiment was to investigate similarities and differences between pairs of English items and their Dutch translations, in terms of both lexical frequency and cognate status (operationalised as phonological similarity). Our expectation was that the lexical frequencies would be rather close (Moscoso del Prado Martín et al. 2004). We also expected that many pairs would show phonological overlap, given the West-Germanic origin of both languages.

Method

Word frequency

Two online corpora were used to obtain word frequencies per million words: the SUBTLEX-US corpus (van Heuven et al. 2014) and the SUBTLEX-NL corpus (Keuleers, Brysbaert, and New 2010) for English and Dutch words, respectively. The US corpus contains 51 million words, and the NL corpus 44 million. The advantage of the SUBTLEX corpora over traditional written corpora is that they are based on film and television subtitles, and thus on spoken language. Frequency estimates based on spoken language seem to be more accurate than estimates based on written language for explaining language processing in children (Brysbaert and New 2009; Keuleers, Brysbaert, and New 2010).

Translations

We translated the items in the PPVT-4 from English to Dutch, making use of the Longman Dictionary of Contemporary English for Advanced Learners (Longman 2012) and the online version of the Van Dale Dutch-English translation dictionary (Albers 2015). One translation was chosen, optimising three criteria: the match with the target picture in the PPVT-4, the closeness of the corresponding meaning of the Dutch word, and the similarity between the frequency of the Dutch translation and the English word. In the PPVT-4, verbs are presented in their -ing form. Such verb forms, expressing ongoing action, are uncommon in Dutch. We therefore used root forms in English and Dutch verb pairs, for example, 'jump' and its Dutch equivalent '*spring*'.

Phonological transcriptions

Phonological transcriptions for all items were retrieved using the Longman Pronunciation Dictionary (Wells 2008). For Dutch, the 'Uitspraakwoordenboek' (Heemskerk and Zonneveld 2000) and the 'Van Dale Middelgroot Woordenboek' dictionary (Albers 2015) were used. We used the X-SAMPA system for these transcriptions.

Objective phonological similarity

We calculated a normalised Levenshtein Distance (LD) in order to determine to what extent word pairs were similar, following Schepens et al. (2013). Because children are presented with the oral and not the written form of the items in the PPVT-4, we chose to focus on the phonological LD. The distance between an English item and its Dutch translation was calculated as the minimum number of insertions, deletions, and substitutions required to go from one to the other in

X-SAMPA notation. All changes were given a weight of 1. For example, for the English item <ankle> and its Dutch translation <enkel>, the phonological difference between English [{Nkl] and Dutch [ENk@l] is 2. Following Schepens et al. (2013) we subtracted the normalised distance from 1 to obtain the phonological similarity:

$$\text{PhonSim} = 1 - \frac{\text{distance}}{\text{length}}.$$

Length was operationalised as the segmental length of the longest word, either the English one or its Dutch translation. The outcome ranges between 0 and 1, where 0 means that there is no overlap between the two strings (completely dissimilar), and 1 means that the strings are identical (completely similar). In case of the example given above, the length of the longest word (the number of segments) was 5, and therefore the phonological distance was calculated as $1 - (2/5) = .60$. If no translation was available (as was the case for three low frequent words), PhonSim was set to 0 (no overlap).

Subjective phonological similarity

In addition to the objective similarity measure, a subjective measure was determined for the first 168 items in the PPVT-4. This allowed us to examine whether the two measures correlate with each other (cf. Potapova, Blumenfeld, and Pruitt-Lord 2016). Participants were 25 native speakers of Dutch (M_{age} = 24.4; SD_{age} = 4.8). They were recruited at Radboud University (Nijmegen, the Netherlands). All participants were volunteers with no known hearing or visual disorders, and gave written consent before taking part in the experiment. They were rewarded with five Euros for their participation.

The similarity rating task was programmed in WebExp2, an application for online experiments developed at Radboud University. Every participant was presented with all 168 item-translation pairs of the PPVT-4, in random order. On each trial, participants were presented with the written English and Dutch form, and the recorded English spoken form that was used in the PPVT-4. Participants were asked to rate the similarity of the English word form and their Dutch translation on a seven-point scale, ranging from *(1) completely different* to *(7) completely similar*. Participants were not presented with the spoken Dutch form to keep the procedure similar to that used in the administration of the PPVT-4. Participants were explicitly asked to pay attention only to the phonetic overlap. The intraclass correlation coefficient was high (*ICC* = .993), showing that the raters strongly agreed in their ratings.

Results and discussion

Word frequencies

Figure 1 shows the average log frequency per million words of the English items and that of their Dutch translations, per set. As expected from how the PPVT-4 was developed, the frequency of the English words declines in the higher sets – although not consistently so. In some sets (e.g. Set 5), the discrepancy between the English and the Dutch frequency is comparatively large, whereas it is minimal in others (e.g. Set 9).

In Figure 2, we plotted the English and Dutch log frequencies against each other for all pairs, to examine how strongly they are related. Although there is an overall positive correlation between English and Dutch frequency ($r = .783$, $p < .001$), several pairs clearly deviate from this tendency. Discrepancies were largest when words had only one or a few meanings in one language and multiple meanings in the other (see Figure 2 for examples).

Phonological similarity

None of the translation pairs were identical, but 15 pairs had a PhonSim of .71 or higher. Figure 3 shows the average PhonSim per set. E.g. the average PhonSim of set 6 is relatively high, whereas that of set 7 is relatively low.

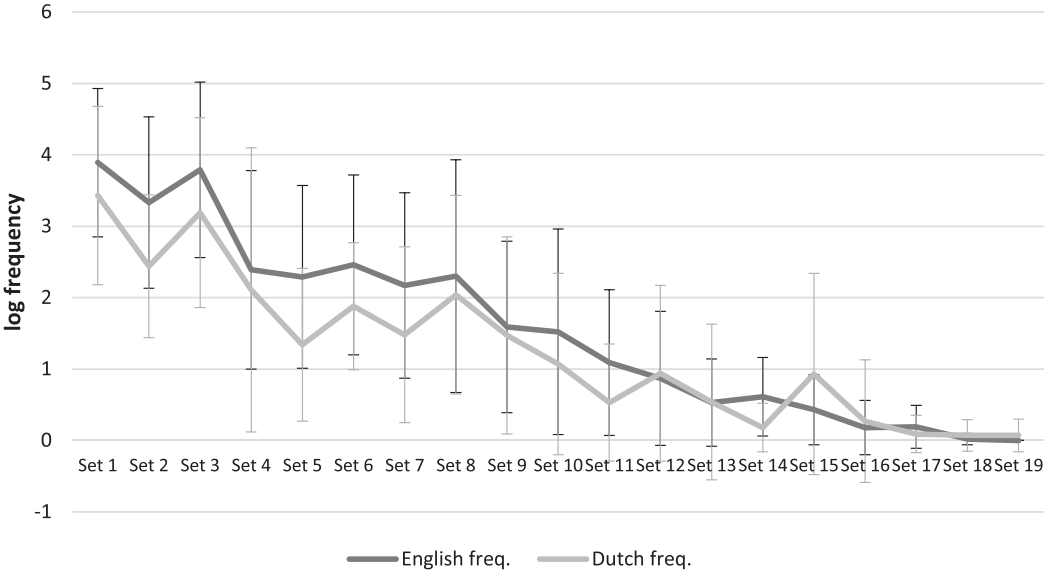

Figure 1. Experiment 1: Average English and Dutch log word frequency per set with their standard errors.

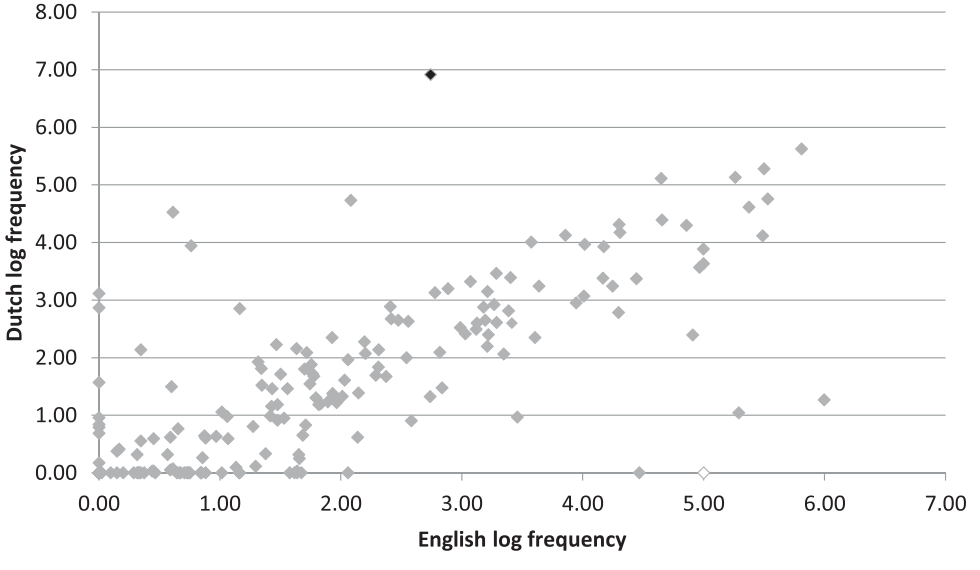

Figure 2. Experiment 1: The relation between English and Dutch log frequency. The word 'net' (depicted by the black dot), for example, has a single predominant meaning in English (something that is made of openwork fabric), and many high-frequency meanings in Dutch (in addition to something made of openwork fabric, it also means 'network', 'tidy', 'decent', and 'exactly'). In the case of 'sort' (depicted by the open dot), which we translated with 'sorteer', it is the other way around.

To assess the validity of the objective measure of form overlap, PhonSim, the correlation between the PhonSim and the subjective similarity judgments was determined. The positive and high correlation found between objective and subjective phonological similarity ($r = .800$, $p < .001$) implies that PhonSim gives a useful, valid measure to identify phonological overlap between words.

In summary, the results of Experiment 1 show that a considerable number of pairs showed substantial phonological overlap. Furthermore, the frequency of the English and Dutch pairs are

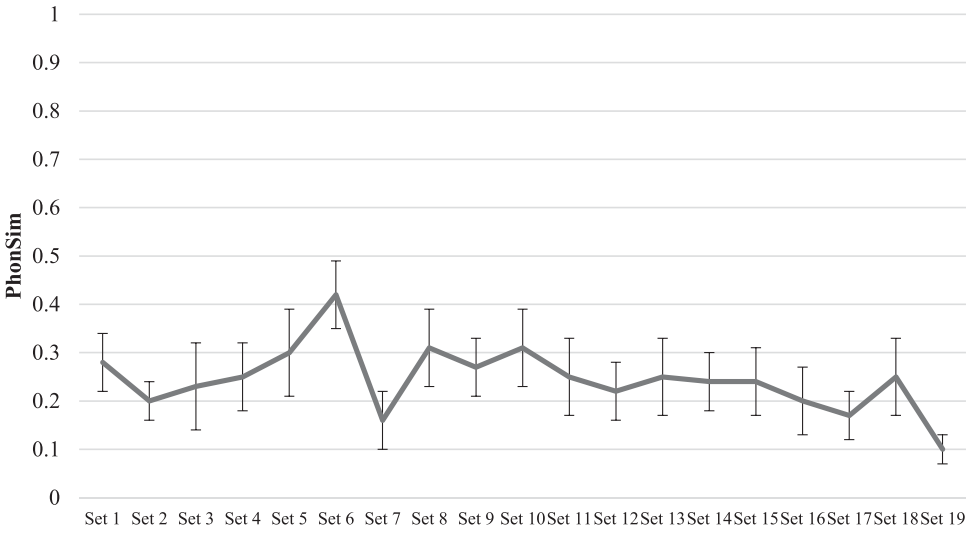

Figure 3. Experiment 1: Average PhonSim per set, and their standard errors.

highly correlated, but crucially not identical, and there is variation in the similarity of the English and Dutch frequency of the translation pairs.

Experiment 2: primary-school pupils

The aim of this experiment was to investigate whether the PPVT-4 is a reliable test when using it with young L2 learners, and to what extent the lexical frequencies and cognate similarity between the English words and their Dutch equivalents found in Experiment 1 could explain primary-school pupils' test performance. Our first hypothesis was that the reliability of the PPVT-4 would increase when administering it to pupils who had more experience with English. Our second hypothesis was that children would perform better on items with more form overlap with their Dutch translations. The third hypothesis was that L1 frequency would positively influence these young pupils' performance.

Method

Participants

Four early-English primary schools that had at least eight years of experience with teaching English participated and four mainstream (i.e. monolingual) schools matched on area (urbanised or rural), neighbourhood (in terms of average income), religious denomination and educational philosophy. Data were collected from 204 typically developing pupils in one of three age groups: 4-to-5 years old (first grade), 8-to-9 years old (fifth grade), or 11-to-12 years old (final grade). An additional 37 pupils participated, but their data were removed from the analyses for one of the following reasons: they had another home language than Dutch ($N = 23$), they did not complete the PPVT ($N = 8$), were learning another language by means of an extracurricular programme ($N = 1$), they were diagnosed with dyslexia ($N = 4$), or they clearly had trouble concentrating during testing ($N = 1$).

The Dutch school system has eight grades, of which the first two are comparable to kindergarten. After the final grade (i.e. grade 8), pupils transfer to secondary school. Pupils are on average 12 years old by then. Pupils from mainstream schools started their English education in the penultimate grade (i.e. at around age 10), and would have had approximately 60 h of English by the end of primary

school. Pupils from early-English schools had had about 320 h of English education by the end of primary school (Jenniskens et al. 2017). In both types of schools, all the rest of the teaching time is in Dutch. Parents gave informed consent for participation and were also asked to complete a questionnaire about out-of-school exposure to English; the response rate was relatively low (37.2%). Table 1 provides the number of girls/boys, exact ages and out-of-school exposure for the type of school and age groups. An independent samples *t*-test revealed that there was no difference between early-English and mainstream pupils in age ($t(202) = 0.208$; $p > .05$) or in out-of-school exposure to English ($t(74) = -1.11$; $p > .05$), but given the low response rate to the questionnaire these results should be interpreted with caution. The PPVT data we present here were previously reported as part of the findings from a larger battery of tests (Goriot et al. 2018).

Instruments

PPVT-4. All children took form A of the Peabody Picture Vocabulary Task – 4th edition (Dunn and Dunn 2007). Pictures were presented on a computer screen. The computer played the accompanying recording, consisting of one word per trial, recorded in a soundproof booth, pronounced in isolation, in clear citation style by a male native speaker of UK English. Administration rules as stated in the manual were followed. Since items in the higher sets (set 15 and higher) were responded to by only a small number of children, too little data were available from these sets to include them in the analyses. We therefore decided to analyse only the first 14 sets, that is, 168 items in total.

Procedure

The PPVT-4 was administered as part of a larger test battery. All children were tested individually in a quiet room at their school. The administration of the complete test battery took place in two sessions, each of which lasted approximately 30 min. The PPVT-4 was administered in the first session. Administration took between 5 and 20 min.

Analysis

First, Cronbach's alpha was computed, to determine the reliabilities for each of the age groups. Cronbach's alpha shows whether all items in one or multiple sets measure the same construct. Second, to examine whether frequency and form overlap played a role in pupils' performances on the PPVT-4 we performed generalised linear mixed-effects model analysis (packages lme4, lmerTest, in platform R, version 3.4.1).

Results

Reliability

We computed pupils' raw score on the PPVT-4, shown in Table 2 for pupils from early-English and mainstream schools separately. For 4-5- and 8-9-year-olds the difference in English vocabulary

Table 1. Experiment 2: number of girls, boys, mean age and mean out-of-school exposure to english per age group.

	Mainstream schools			Early-English schools		
	4-5 year-olds	8-9 year-olds	11-12 year-olds	4-5 year-olds	8-9 year-olds	11-12 year-olds
N	38	34	26	40	38	28
Girls (N)	21	17	16	23	19	19
Boys (N)	17	17	10	17	19	9
Age (*M, SD*)	4.90 (0.27)	9.08 (0.40)	12.15 (0.53)	4.81 (0.35)	8.95 (0.41)	11.97 (0.38)
Out-of-school exposure to English in hours per week *(M, SD)*	5.09 (6.08)	9.27 (6.35)	16.10 (14.11)	9.14 (7.97)	7.44 (6.41)	22.5 (16.43)

Table 2. Experiment 2: vocabulary scores by age group and type of education.

		4-5 year-olds		8-9 year-olds		11-12 year-olds	
		Mainstream	Early-English	Mainstream	Early-English	Mainstream	Early-English
N		38	40	33	38	22	21
PPVT-4 raw score	*M (SD)*	13.4 (8.3)	18.2 (14.8)	53.5 (22.7)	56.1 (16.8)	84.1 (21.3)	101.1 (21.0)

Table 3. Experiment 2: Cronbach's alpha for the different age groups and sets.

	4-5 year-olds		8-9 year-olds		11-12 year-olds	
	Alpha	Cumulative N (pupils)	Alpha	Cumulative N (pupils)	Alpha	Cumulative N (pupils)
Set 1	−4.085	5	-	-	-	-
Set 1-2	.431	31	-	-	-	-
Set 1-3	.784	46	-	1	-	-
Set 1-4	.834	66	.875	3	-	-
Set 1-5	.863	73	.888	11	.904	2
Set 1-6	.886	75	.846	15	.904	2
Set 1-7	.916	78	.807	38	.904	2
Set 1-8			.816	41	.903	4
Set 1-9			.873	48	.879	5
Set 1-10			.909	56	.956	9
Set 1-11			.944	66	.940	23
Set 1-12			.948	69	.933	29
Set 1-13			.950	70	.955	35
Set 1-14			.954	71	.962	43

scores between mainstream and early-English pupils is small. For 11-12-year-olds, the difference is larger.

Table 3 shows Cronbach's alpha for the different age groups separately. We calculated Cronbach's alpha over a cumulative number of items, starting with the pupils that only completed the first set of items, thereafter also including pupils that completed the second set, and so on. In this way, we can see what the reliability is at lower proficiency levels, and whether reliability systematically increases when more sets are included. Indeed, as predicted, the higher reliability scores are found in the older age groups, particularly in the 11-12-year-olds. The largest increases in reliability values are found when including more sets for the 4-5-year-olds. Below set 4, the reliability scores are low (<.600) to medium (<.800), meaning that the PPVT-4 does not produce reliable scores ($\alpha \geq .900$; McNamara 2000) in the lowest scoring group of L2 learners.

Table 4 shows that percentages correct for all age groups are positively correlated with cognate status as measured by PhonSim: the closer the English word and its Dutch translation are, the larger the percentage correct, and thus the easier the item. Frequencies in Dutch and English are also positively related to percentages correct, except in the youngest age group.

Table 4. Experiment 2. Correlations between phonological similarity, frequencies, and percentage correct.

	1	2	3	4	5	6
1. PhonSim	1.					
2. Freq. EN	.037	1.				
3. Freq. NL	.094	.754**	1.			
4. Overall Percentage correct	.569**	.139	.233**	1.		
5. Percentage correct age 4-5	.490**	-.047	.111	.724**	1.	
6. Percentage correct age 8-9	.476**	.187*	.251**	.825**	.696**	1.
7. Percentage correct age 11-12	.475**	.346**	.377**	.907**	.493**	.757**

*$p < .015$, **$p < .01$.

Table 5. Experiment 2: parameter estimates from the model for primary-school pupils' vocabulary scores.

	Fixed effects		
Parameters	Estimate	SE	*Z*-value
Intercept	−1.83	0.21	−8.80***
8-9 year-olds[a]	1.34	0.18	7.60***
11-12 year-olds[a]	2.07	0.19	10.80***
Type of Education[b]	0.33	0.18	1.89(*)
English frequency	−0.14	0.14	−1.01
Dutch frequency	0.32	0.15	2.17*
PhonSim	1.45	0.59	2.45*
8-9 year-olds × Type of Education	−0.32	0.21	−1.52
11-12 year-olds × Type of Education	0.16	0.23	0.72
8- 9 year-olds × En. Freq.	0.21	0.08	2.60**
11-12 year-olds × En. Freq.	0.39	0.08	4.63***
8-9 year-olds × Dutch Freq.	0.04	0.08	0.49
11-12 year-olds × Dutch Freq.	0.06	0.08	0.71
8-9 year-olds × PhonSim	2.34	0.32	7.24***
11-12 year-olds × PhonSim	3.01	0.36	8.26***

[a]The 4-5 year-olds are the reference group, [b]Mainstream education is the reference group.
(*)$p < .06$, *$p < .05$, **$p < .01$, ***$p < .001$.

Generalised linear mixed effects model analysis

We conducted a generalised linear mixed-effects model analysis with binomial responses (incorrect coded as 0 and correct coded as 1) as dependent variable, Item number, Subject, and School as random effects, and Type of Education, Age Group, English Frequency, Dutch Frequency, PhonSim, and the interactions between Age Group and all other variables as fixed effects. Continuous variables (English Frequency, Dutch Frequency, and PhonSim) were mean-centered. The results, displayed in Table 5, show that there is a main effect of Age Group: the 8-9- and 11-12-year-olds performed better than the 4-5-year-olds. There is also a positive main effect of Dutch frequency. Early-English pupils tended to perform better on the PPVT-4 than mainstream pupils, resulting in a marginally significant main effect of Type of Education ($p = .058$). The effects of English frequency and PhonSim were different for 8-9-year-olds and 11-12-year-olds than for 4-5-year-olds (see Figures 4 and 5, respectively), as shown by the interaction effects between Frequency and Age Group, and PhonSim and Age Group.

To gain insight in the effects of frequencies and cognate status on the performance of pupils of different ages, we conducted follow-up analyses for the three age groups separately. For each of the age groups, we fitted the same model as reported previously (but without Age Group as a predictor). The results are shown in Table 6. Type of Education only showed a significant main effect for 11-12-year-olds: early-English pupils performed better than mainstream pupils. Similarly, English frequency was a significant and positive predictor of correct responses in the highest age group only. Dutch Frequency on the other hand was a positive and significant predictor of correct responses in the 8-9-year-olds and the 11-12-year-olds; in the youngest group, Dutch Frequency showed a trend towards significance ($p = .057$). Phonological distance was a positive and significant predictor in all three age groups.

Discussion

Confirming our first hypothesis, this experiment showed that the reliability of the PPVT-4 was quite low for pupils who completed only the first three sets (59% of the 4-5-year-olds), but became increasingly higher for pupils who reached the higher sets. Second, we found that, confirming our second hypothesis, pupils performed better on items that sounded more similar to their Dutch translations. Third, we found that pupils performed better on items that were more frequent in Dutch, whereas English frequency was a positive predictor of 11-12-year-olds' performance only. These findings partly confirm our third hypothesis, namely that for young pupils Dutch frequency is an important

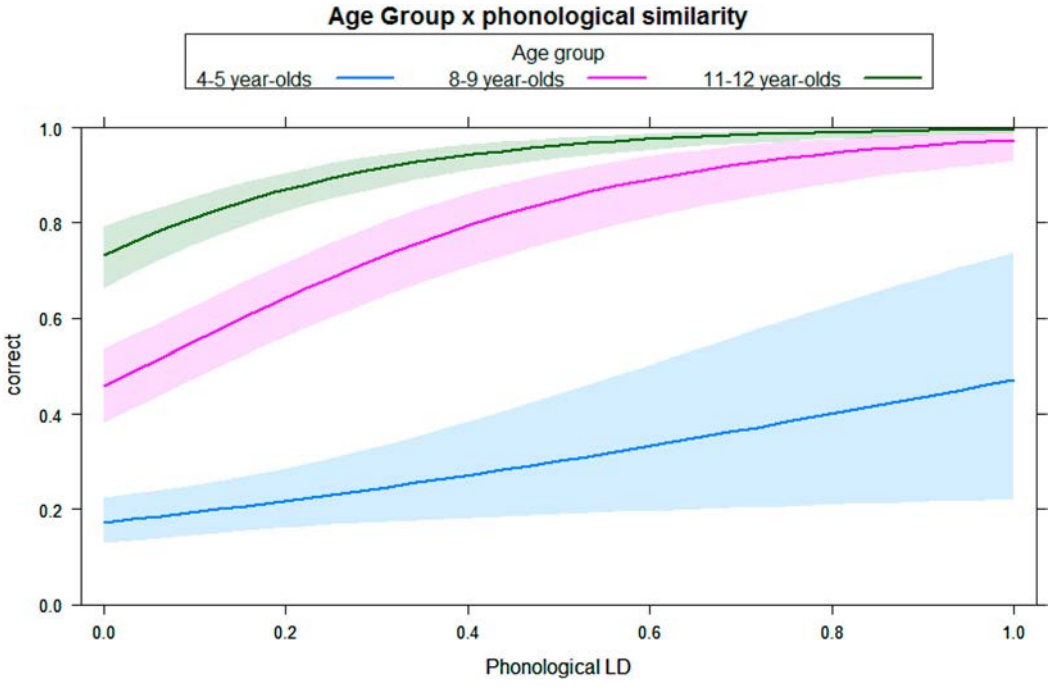

Figure 4. Experiment 2: Relation between PhonSim and PPVT scores for the different age groups.

predictor of performance. Contrary to the hypothesis, for 4-5-year-olds, Dutch frequency only played a marginally significant role. We will elaborate on this finding in the general discussion.

Given that most of the children did not respond to any items from set 15 onwards, it was not possible to draw any conclusions about the performance on later items in the test, or on the possible frequency and similarity effects of those items. To investigate what the role of L1 effects in L2 vocabulary testing is in older children who have more experience with the English language, as well as to investigate the quality of the remaining items in the PPVT-4, we collected data from secondary-school pupils and ran the same analyses.

Experiment 3: secondary-school pupils

We replicated the primary-school experiment (Experiment 2) with secondary-school pupils. Based on the results of Experiment 2, our first hypothesis was that the reliability of the PPVT-4 would be high when administering it to secondary-school pupils. The second hypothesis was that, again, pupils' performance would be positively influenced by phonological similarities between item-translation pairs. Finally, our third hypothesis was that, contrary to primary-school pupils, the English frequency of the items would positively influence the performance of secondary-school pupils, whereas the influence of Dutch frequency would decrease.

Method

Participants

One school participated that had provided both a bilingual and a mainstream (i.e. monolingual) curriculum for six years. Data were collected from 152 pupils, who where in the first year (12-13-year-olds; 14 female, 21 male), second year (13-14-year-olds; 27 female, 26 male), or third year (14-15-year-olds; 35 female, 32 male). All pupils were in the pre-university track (Dutch: 'VWO'), which

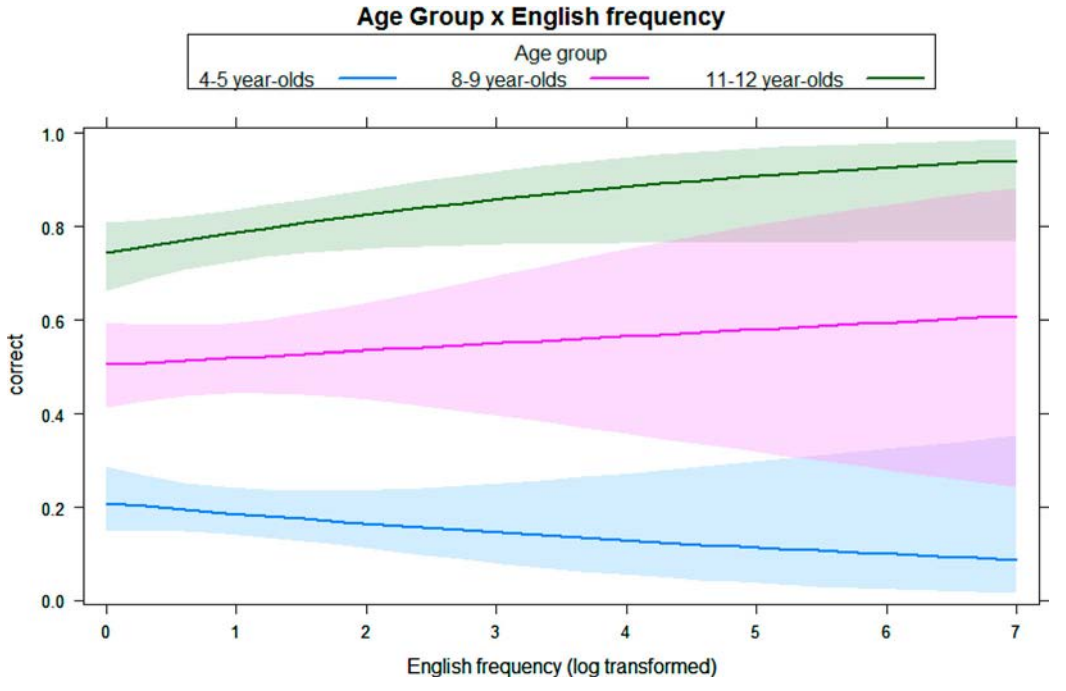

Figure 5. Experiment 2: Relation between English frequency and vocabulary scores, for the different age groups.

Table 6. Experiment 2: parameter estimates for the models for primary-school pupils from different age groups.

	4-5 year-olds Fixed effects			8-9 year-olds Fixed effects			11-12 year-olds Fixed effects		
Parameters	Estimate	SE	*Z*-value	Estimate	SE	*Z*-value	Estimate	SE	*Z*-value
Intercept	−1.28	0.27	−4.39***	−0.37	0.17	−2.26*	0.29	0.25	1.67
Type of Education	0.29	0.24	1.23	0.01	0.12	0.12	0.56	0.27	2.12*
English frequency	−0.12	0.14	−0.91	0.04	0.13	0.31	0.30	0.15	1.97*
Dutch frequency	0.24	0.13	1.90(*)	0.34	0.13	2.55*	10.44	0.16	2.71**
PhonSim	2.84	0.58	4.88***	3.68	0.54	6.85***	4.41	0.65	6.75***
AIC		3546.6			7214.5			5692.9	

(*)$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

upon completion gives admission to university education (the Dutch educational system is selective, with three secondary-school tracks). Pupils had been following either the mainstream or the bilingual curriculum since the start of secondary school (see Table 7): those in the mainstream curriculum followed their lessons in Dutch, but had lessons on English as a foreign language for 150 (12-13-year-olds) or 120 min per week (13-14- and 14-15-year-olds). Pupils in the bilingual curriculum received half of their subject lessons (approximately 800 min per week) in English. All pupils gave informed consent for participation.

Table 7. Experiment 3: number of pupils in the mainstream and the bilingual curriculum.

Year	Mainstream curriculum	Bilingual curriculum
1	15	20
2	2	51
3	18	46
Total	35	117

Procedure

Pupils were tested individually in a quiet room in the school during school time. All pupils performed the PPVT-4, which was part of a larger test battery (but a different test battery than in Experiment 2). The administration procedure was slightly different from the procedure in Experiment 2: all pupils started with the first set instead of with the age-appropriate set. Contrary to Experiment 2, for part of the pupils, printed pictures were shown, and words were read by the experimenter. The other pupils were presented with a computerised version of the PPVT-4 similar to the one used in Experiment 2. Similar to Experiment 2, testing stopped when pupils made eight or more errors, as indicated in the manual. Administration of the PPVT-4 took approximately 20 min.

Analysis

In a similar fashion to Experiment 2, we examined Cronbach's alpha and the percentage correct to investigate the reliability of the different sets, and the difficulty of the items. Again, a generalised linear mixed-effects analysis was conducted in order to investigate effects of L1 and of Type of Education on performance.

Results

Figure 6 shows the scores on the PPVT-4, both for primary-school (Experiment 2) and secondary-school pupils (Experiment 3). It shows that, descriptively, early-English pupils have higher scores than mainstream pupils, and older pupils have higher scores than younger pupils. Table 8 shows that the reliability of the sets is excellent in all three secondary-school years.

Figure 7 shows the percentage correct per set, for each age group, again both for primary-school (Experiment 2) and secondary-school pupils (Experiment 3). For the secondary-school pupils (12- to 15-year-olds), the first sets seem to be easy with performance at (or near) ceiling. The percentage correct declines steeply in the later sets.

We correlated the percentages correct for the different year groups with PhonSim, the Dutch frequencies, and the English frequencies (Table 9). The results were highly similar to the results in the

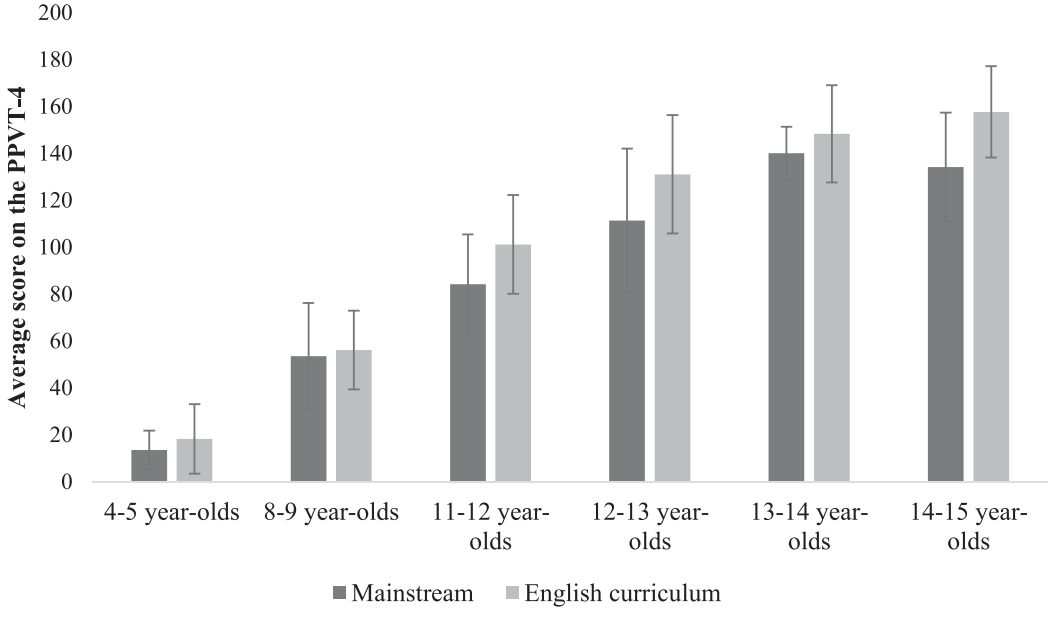

Figure 6. Experiments 2 and 3: Average scores (raw number) on the PPVT-4 with standard deviations. Note: In the 13-14 year-olds, the mainstream group consists of only two pupils.

Table 8. Experiment 3: Cronbach's alpha for the different age groups and sets.

	12-13 year-olds (Year 1)		13-14 year-olds (Year 2)		14-15 year-olds (Year 3)	
Set	Alpha	Cumulative N (pupils)	Alpha	Cumulative N (pupils)	Alpha	Cumulative N (pupils)
Set 1	-	-	-	-	-	-
Set 1-2	-	-	-	-	-	-
Set 1-3	-	-	-	-	-	-
Set 1-4	-	-	-	-	-	-
Set 1-5	-	-	-	-	-	-
Set 1-6	-	-	-	-	-	-
Set 1-7	-	-	-	-	-	-
Set 1-8	-	-	-	-	-	-
Set 1-9	-	-	-	-	-	-
Set 1-10	.968	2	-	-	-	-
Set 1-11	.956	9	.942	4	.943	3
Set 1-12	.952	10	.942	4	.913	4
Set 1-13	.946	15	.942	6	.940	6
Set 1-14	.942	21	.950	9	.932	10
Set 1-15	.963	28	.950	30	.943	32
Set 1-16	.964	29	.942	40	.942	42
Set 1-17	.971	33	.946	49	.955	54
Set 1-18	.974	35	.950	51	.960	61
Set 1-19			.956	53	.967	64

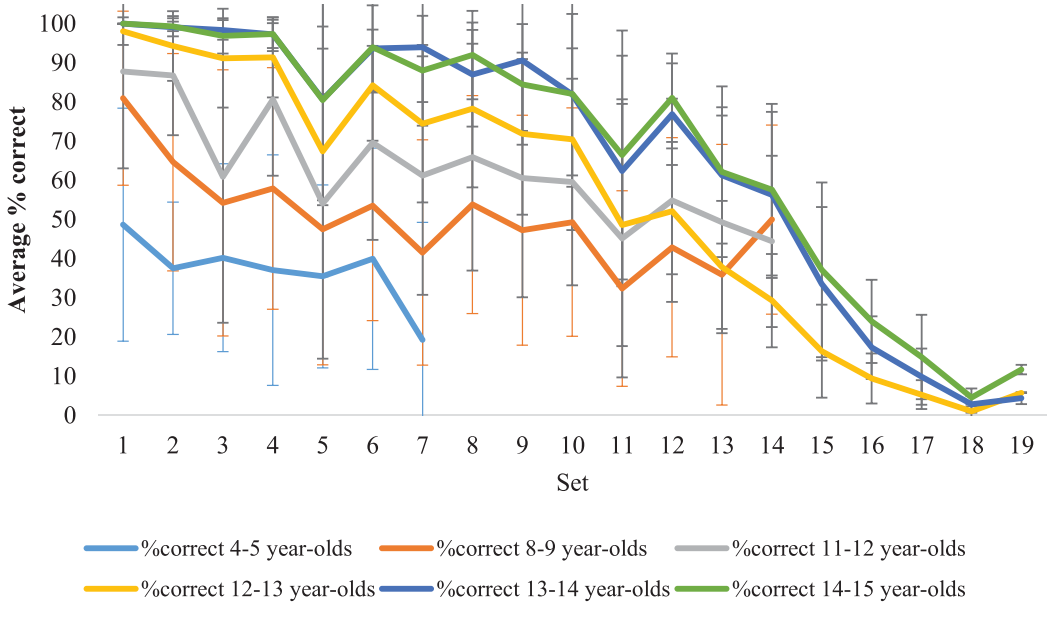

Figure 7. Experiments 2 and 3: Percentages correct per set, for each age group (with SDs).

younger children: Percentage correct correlated significantly with PhonSim and with frequencies in English and in Dutch.

Generalised linear mixed-effects analysis

We performed a general linear mixed-effects analysis on pupils' performance (0 as incorrect and 1 as correct) on the items of the PPVT-4, with Age Group, Type of Education, English Frequency, Dutch Frequency, PhonSim, and the interactions between Age Group and the other variables. Random slopes were included at the Subject and Item level. Pupils in the 13-14-year-old age group were

Table 9. Experiment 3. Correlations between phonological similarity, frequencies, and percentage correct.

	1	2	3	4	5	6
1. PhonSim	1					
2. Freq. EN	.124	1				
3. Freq. NL	.132*	.759***	1			
4. Overall percentage correct	.236***	.634***	.543***	1		
5. Percentage correct 12-13 year-olds	.256***	.661***	.580***	.970***	1	
6. Percentage correct 13-14 year-olds	.210**	.615***	.534***	.998***	.949***	1
7. Percentage correct 14-15 year-olds	.227***	.601***	.520***	.982***	.940***	.986***

*$p < .046$, **$p < .01$, ***$p < .001$.

Table 10. Experiment 3. Parameter estimates for the model for secondary-school pupils' vocabulary scores.

Parameters	Fixed effects		
	Estimate	SE	*Z*-value
Intercept	0.02	0.14	0.11
14-15 year-olds[a]	0.85	0.07	11.68***
Type of Education[b]	0.68	0.07	9.79***
English frequency	1.10	0.14	7.98***
Dutch frequency	0.20	0.15	1.37
PhonSim	4.70	0.59	7.94***
14-15 year-olds × Type of Education	0.41	0.09	4.47***
14-15 year-olds × En. Freq.	0.03	0.05	0.60
14-15 year-olds × Dutch Freq.	0.07	0.05	1.28
14-15 year-olds × PhonSim	0.10	0.21	0.47
AIC	14718.2		

[a]The 12-13 year-olds are the reference category, [b]Mainstream education is the reference category.
***$p < .001$.

left out of the analysis, because of the low number of participants ($N = 2$) in the mainstream curriculum. The results, in Table 10, show that there are main effects of all independent variables except for Dutch Frequency. Older pupils perform better on the items in the PPVT-4 than younger pupils. Pupils perform better on items that are more frequent in English, and on items that sound more similar to their Dutch translations. Pupils in the bilingual curriculum outperformed pupils in the mainstream curriculum; the differences between pupils from the mainstream and the bilingual curriculum were smaller in 12-13- than in 14-15-year-olds (see Figure 8), resulting in a significant interaction.

To investigate the effects of the independent variables on the performance of pupils in the different age groups separately, we performed the same analysis again but now for each age group separately (without Age Group as a predictor). The outcomes are shown in Table 11. For 14-15-year-old pupils, there was a main effect of Type of Education in favour of pupils in the bilingual curriculum. For the 12-13-year-olds this effect was only marginally significant ($p = .057$). In both age groups, English frequency and PhonSim were significant and positive predictors of correct responses. Dutch frequency was never significant, although it seemed that the oldest pupils performed better on items that were more frequent in Dutch ($p = .093$).

Discussion

Confirming our first hypothesis that the reliability of the PPVT-4 would be high when administering it to secondary-school pupils who have more experience with English, we found that the reliability of the PPVT-4 was always ≥.900. Our second hypothesis, that pupils would perform better on English items that are phonologically closer to their Dutch translations, was also confirmed. Our third hypothesis was that English frequency would positively influence secondary-school pupils' performance, and that, contrary to primary-school pupils, the role of Dutch frequency would be decreasing. The results

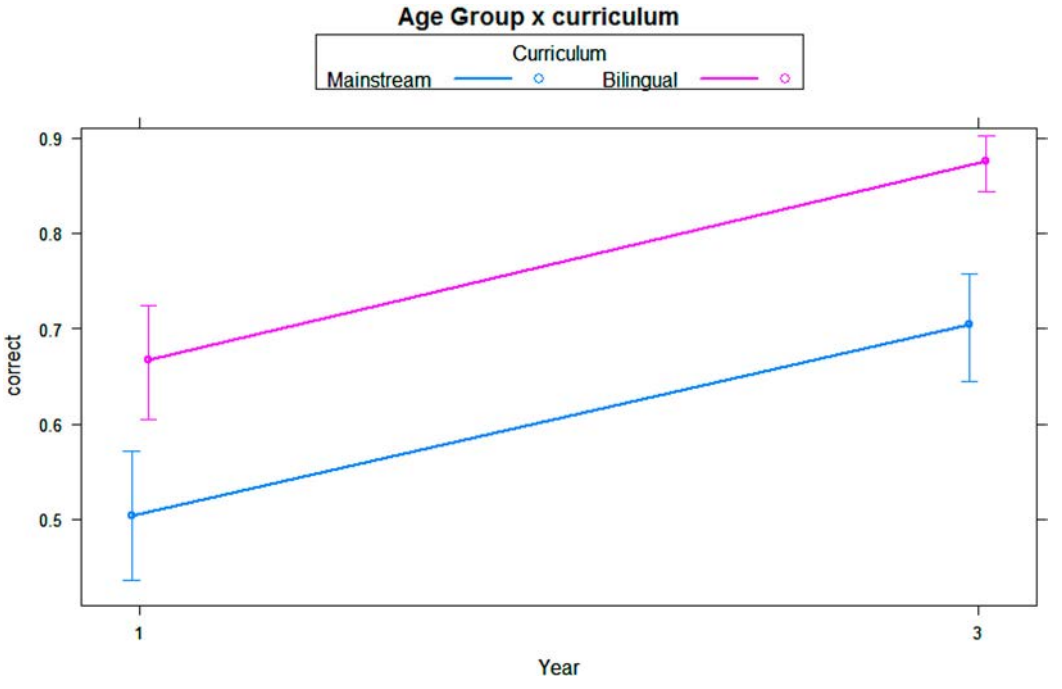

Figure 8. Experiment 3: The relation between age and correct scores, for pupils from the two curricula.

Table 11. Experiment 3. Parameter estimates for the models for secondary-school pupils from different age groups.

	12-13 year-olds Fixed effects			14-15 year-olds Fixed effects		
Parameters	Estimate	SE	*Z*-value	Estimate	SE	*Z*-value
Intercept	0.05	0.35	0.16	1.04	0.32	3.21**
Type of Education	0.79	0.42	1.90(*)	1.33	0.33	4.05***
English frequency	1.33	0.16	8.24***	1.36	0.18	7.73***
Dutch frequency	0.24	0.17	1.38	0.32	0.19	1.68(*)
PhonSim	5.52	0.69	7.95***	5.64	0.76	7.42***
AIC		5188.5			7957.7	

Note: 12–13 year–olds: (*)$p = .057$, 14–15 year–olds: (*)$p = .093$, **$p = .001$.
(*)$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

confirmed the hypothesis in the sense that pupils performed better on items that were more frequent in English, while Dutch frequency was not a significant predictor of pupils' performance. Contrary to our expectations Dutch frequency did however show a trend towards significance in the oldest age group. We will come back to this finding in the general discussion.

General Discussion

The first aim of this study was to examine the reliability of the PPVT-4 as a tool for measuring L2 vocabulary knowledge of Dutch learners of English. The second aim was to examine whether L1 characteristics, in particular lexical frequency and cognate similarity (operationalised by phonological similarity) of the item-translation pairs, could affect scores on the PPVT-4 when using it to measure L2 vocabulary. We investigated this question by administering the PPVT-4 to primary-school and secondary-school pupils of different age groups who were learning English as an L2. We investigated the technical quality of the test, and the characteristics of the individual items.

We also examined the relation between L1 and L2 word frequency and cognate status on the one hand, and primary and secondary-school pupils' vocabulary scores on the other.

Because English and Dutch are both Germanic languages, we expected that there would be relatively strong form overlap in the English-Dutch item-translation pairs. We also expected that Dutch and English lexical frequency of the PPVT-4 items would be closely related to each other (Moscoso del Prado Martín et al. 2004). The results of Experiment 1 showed that this was indeed the case: English and Dutch lexical frequency correlated positively and highly with each other, but were not identical. Furthermore, there was ample variation in the similarity of the two items' frequencies across translation pairs. There was also substantial phonological similarity between the English items in the PPVT-4 and their Dutch translations, thereby confirming previous results that showed that English and Dutch share many cognates (Schepens et al. 2013).

Our first hypothesis was that the PPVT-4 would be a more reliable vocabulary measure when administering it to pupils who have more experience with English then when administering it to less experienced pupils. The latter group was expected to complete only the first few sets of the PPVT-4, and we expected that the test would not reliably differentiate between pupils' scores based on a limited number of items. Indeed, for 4-5-year-olds who only made it to the first three sets, Cronbach's alpha was low ($\alpha < .600$) to medium ($\alpha < .800$). The test reached an acceptable reliability level ($\alpha > .900$; McNamara 2000) when pupils completed more sets.

Our second hypothesis was that phonological similarity between the English words and their Dutch translations would positively influence pupils scores on the PPVT-4, as a cognate effect had already been shown for Spanish children (Potapova, Blumenfeld, and Pruitt-Lord 2016). Moreover, we expected that this effect would be larger for older pupils, as previous research had shown that older Dutch-Frisian bilingual children are better at recognising overlap between words in two languages than younger children (Bosma et al. 2016). Extending previous research (Bosma et al. 2016; Pérez, Peña, and Bedore 2010; Potapova, Blumenfeld, and Pruitt-Lord 2016), we showed that both for younger and older pupils, phonological similarity between English and Dutch words was a positive predictor of pupils' performance. As expected, this effect was larger for older pupils than for younger pupils. This suggests that older pupils may be more able than younger pupils to make use of phonological similarities between item-translation pairs for the comprehension of L2 words.

We assessed phonological rather than orthographic similarity between the English and Dutch words. This means that the extent to which orthographic similarity predicts pupils' L2 vocabulary scores remains unknown. We believe that our choice to investigate phonological similarity is theoretically well-grounded, for two reasons. The first is that in the PPVT-4, children are typically presented with the words orally. The second reason is that pupils are probably more familiar with the oral than with the written form of words, if they are familiar with the written forms at all: the focus of English lessons in primary school is mostly on oral skills (Jenniskens et al. 2017; Thijs et al. 2011), and, in secondary school, pupils receive content and language integrated learning (CLIL) contexts in the bilingual programme, in which the focus is on meaning and not form (Dalton-Puffer 2011). In other words, whilst we cannot rule out a possible influence of orthographic similarity in addition to phonological similarity, it is unlikely that the former would be greater than the latter.

Our third hypothesis was that the influence of L1 (Dutch) frequency would decrease in older, more experienced pupils, while the influence of L2 (English) frequency on pupils' scores would increase. Even within early-English educational programmes, primary-school pupils are generally exposed to English for maximally one hour per week (Jenniskens et al. 2017). Since pupils would not have had enough English exposure for English frequency to matter a great deal, we expected that pupils would mainly rely on their knowledge of the items in Dutch, and thus the Dutch frequency of the words would be a better predictor of their vocabulary scores. Secondary-school pupils would be more exposed to English, both inside and outside the school (Lindgren and Muñoz 2013), and hence English frequency may play a more important role in this group. Our hypothesis was confirmed: Dutch frequency was positively related to primary-school pupils' performance on the PPVT-4, whereas for secondary-school pupils it was not. English frequency was only a significant

predictor of scores for the oldest (i.e. 11-12-year-old) primary-school pupils and for the pupils in secondary school. Wood and Pena (2015) showed that the relation between children's errors and difficulty level of the items in the PPVT-4 (as measured by the ordering of the items), was stronger for English L1 children than for Spanish children learning English as an L2. We extended these findings by showing that English L2 learners' performance on the PPVT seems to depend, in the least experienced group, on the items' frequency in their L1, rather than on the frequency in the L2. For 4-5-year-olds, who are in this study also the least experienced, Dutch frequency was however only marginally significant. It is possible that because these pupils completed relatively few items, there may not have been enough variation in the Dutch lexical frequencies to show a significant relation with their performance on the PPVT in English. Furthermore, in the oldest, and in this study the most experienced group, pupils seemed to perform better on items that had a higher frequency in Dutch, although this relation was not significant. It may be that as these children encountered the low-frequent words at the end of the test, their performance depended on whether they knew the word in Dutch in the first place.

We also asked whether pupils following an English curriculum at school would differ in their performance on the PPVT-4 from their peers who followed the mainstream curriculum. We indeed found a difference in favour of the pupils in the English curriculum, but only in the older pupils. For the 4-5-year-olds and 8-9-year-olds, there was no significant difference between pupils from the two types of education. Several reasons could account for the absence of this difference. The first reason could be that there really are no differences between the groups. Although previous research has shown that early-English pupils in kindergarten (Unsworth et al. 2015) and in the final grade of primary school (de Graaff 2015) outperformed their peers from mainstream schools as a group, it has also been shown that the performance scores of the groups overlapped, such that individual pupils from mainstream schools outperformed individuals from early-English schools on English proficiency tests (de Graaff 2015). Note that in the latter study, pupils were not tested on vocabulary but on spelling, listening and reading skills, and on 'use of English'. Furthermore, those two previous studies have shown that, besides early-English versus mainstream education, the development of English is related to other factors, such as the English proficiency level of the teacher, the amount of English input in school (Unsworth et al. 2015), and out-of-school-exposure to English (de Graaff 2015). It may be the case that the early-English participants in our study got less than one hour of English input in school or that they were educated by a teacher with a moderate proficiency level of English, two factors that have been shown to result in lower vocabulary scores compared to pupils who receive more input in English, or who are educated by a (near-)native speaker of English (Unsworth et al. 2015). In addition, the response rate to the parental questionnaire was very low, making it hard to draw conclusions about out-of-school exposure to English. It may be the case that the mainstream pupils received more out-of-school exposure to English than the early-English pupils, which may have compensated for the lack of English instruction at school.

The second reason for the absence of a significant difference between pupils from the mainstream and early-English curriculum might be the low reliability of the PPVT-4, especially in the 4-5 year olds. In previous research, differences between early-English and mainstream kindergartners were small (Unsworth et al. 2015). It may thus be that there are in fact differences between the pupils of the two types of education in our study, too, but that these subtle differences do not always show when administering the PPVT-4 as outlined in the test manual. We therefore suggest, when administering the PPVT, to start with an earlier set than the age-appropriate set.

Implications for research and practice

Our study has shown that L1 lexical frequency and cognate similarity (operationalised as phonological similarity) influenced Dutch pupils' performance on the PPVT-4 as a measure of L2 English vocabulary. This is reminiscent of findings with adults that show that acquiring a new language is easier when that language is close to your native language (Schepens, van der Slik, and van Hout 2016), and that

linguistic similarity helps learners to derive the meaning of foreign words (Pérez, Peña, and Bedore 2010; Potapova, Blumenfeld, and Pruitt-Lord 2016). Related languages overlap by definition, and are thus very likely to contain translation equivalents that also share aspects of their form. Previous research with adults has already shown that L2 learners with different mother tongues obtain significantly different scores, depending on the number of cognates between their L1 and the words in the PPVT (Leśniewska, Pichette, and Béland 2018). Dutch and English are relatively close to each other, having a relatively large number of cognates (Lindgren and Muñoz 2013; Schepens et al. 2013). Since effects of L1-L2 similarities are different for different L1s, researchers should therefore be cautious in comparing receptive English vocabulary knowledge of children across L1s, as is sometimes done (Enever et al. 2011; Lindgren and Muñoz 2013; Steinlen and Piske 2013). Our findings suggest that Dutch children will perform better on the PPVT-4 than children with a non-Germanic language as mother tongue because of the higher proportion of cognate items for the Dutch children.

When investigating the question whether children following an English programme at school have better knowledge of English vocabulary than pupils who do not follow such a programme, researchers should ideally make use of a curriculum-independent vocabulary test that is able to capture English vocabulary knowledge that pupils learned in school. Since such a test does not exist, in any case not for the Dutch context, the PPVT-4 is often used to answer the question whether early-English pupils have a better developed English vocabulary than pupils enrolled in mainstream schools. We investigated whether it is suitable to use the PPVT-4 to answer this question. As many as 31 out of 78 (40%) 4-5-year-old participants completed no more than the first two sets of the PPVT-4. The reliability of these two sets was very low, which suggests that the PPVT-4 may not be suitable for use with unexperienced (or younger) L2 learners. The higher values for the older groups suggest that the test is more reliable when using it with more experienced (or older) L2 learners. Nevertheless, effects that were unaccounted for in the design of the test, such as cognate status and L1 word frequency, still play an important role in these groups. Researchers should thus be cautious in interpreting the results.

This study had a cross-sectional design. Including pupils from six different age groups has provided us with insight in what factors may play a role in L2 vocabulary testing at different ages. We cannot be certain, however, about the (causal) developmental pattern of the L1 and L2 factors at play. A longitudinal experiment would provide more insight in the relation between the extent to which Dutch and English lexical frequency and cognate status predict performance on the PPVT-4 as pupils grow older. Further, including a group of very young learners from a bilingual programme may reveal whether the test becomes a more reliable tool to test English vocabulary in young learners when they have more knowledge of English.

Conclusion

We investigated the use of the PPVT-4 as a measure of receptive English vocabulary in L2 learners. We found that in young primary-school pupils, the frequency of the Dutch translations of the English items as opposed to the frequency of the English test words themselves positively related to their performance on the test. Both primary and secondary-school pupils performed better on English items that were phonetically closer to their Dutch translations. These findings indicate that pupils' L1 plays a role when assessing vocabulary in the L2. Researchers should be aware of these influences, especially when comparing pupils with different mother tongues. Nevertheless, the PPVT-4 seems to be a suitable curriculum-independent instrument for the relative ranking on L2 English vocabulary size of more experienced L2 learners with the same mother tongue.

Acknowledgements

We thank Nuffic (formerly EP-Nuffic) for financial support for Experiment 2, and Varendonck College, Asten (and Vereniging Ons Middelbaar Onderwijs) for financial support for Experiment 3. We also thank Max Planck Institute for

Psycholinguistics for lending us the PPVT-4. We are particularly grateful to the students who helped with data collection. During the period of carrying out this research, Mirjam Broersma and Sharon Unsworth were supported by a Vidi Grant from NWO (the Netherlands Organisation for Scientific Research).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Nuffic (formerly EP-Nuffic).

Notes on contributors

Claire Goriot is a PhD candidate at the Centre for Language Studies of Radboud University, Nijmegen. Her research focuses on educational processes, (language) learning, and child bilingualism.

Roeland van Hout is a professor of Applied and Variational Linguistics at the Center for Language Studies of the Radboud University Nijmegen. His work focuses on language variation and change and second language acquisition, from an interdisciplinary perspective (sociology, linguistics, psychology). He also publishes on the application of statistics in language research.

Mirjam Broersma received a doctoral degree in Social Sciences from the Radboud University in 2005, as well as a Max Planck Society Otto Hahn Medal for outstanding scientific achievements, for her dissertation Phonetic and lexical processing in a second language. In 2006, she started the project Relearning a lost language: Speech perception in Korean by adoptees, funded by a three-year Veni career development grant from the Netherlands Organisation for Scientific Research (NWO). She currently leads the project We learn from our mistakes – or do we? Towards more efficient use of talking and listening experience in a second language, funded by a five-year Vidi career development grant and an Aspasia individual grant from NOW.

Vanessa Lobo received her doctorate from Radboud University in 2013. She is interested in second language acquisition in an educational context and is herself an experienced teacher of English. She has always combined research with teaching in secondary education.

James M. McQueen is Professor of Speech and Learning at Radboud University, Nijmegen. His research focusses on learning and processing in spoken language: How do listeners learn the sounds and words of their native and non-native languages, and how do they recognize them? His research on speech learning concerns initial acquisition processes and ongoing processes of perceptual adaptation. His research on speech processing addresses core computational problems (such as the variability and segmentation problems).

Sharon Unsworth (Ph.D. 2005, Utrecht University) is an Associate Professor of Second Language Acquisition in the department of English Language and Culture, the department of Linguistics and the Centre for Language Studies at Radboud University, Nijmegen, the Netherlands. Her research focuses on the language development of bilingual children and children acquiring a second language in early childhood. Recent publications address the role of input and age, cross-linguistic influence and the operationalisation of language dominance.

ORCID

Claire Goriot http://orcid.org/0000-0001-9088-3908

References

Admiraal, W., G. Westhoff, and K. de Bot. 2006. "Evaluation of Bilingual Secondary Education in the Netherlands: Students' Language Proficiency in English." *Educational Research and Evaluation* 12 (1): 75–93. doi:10.1080/13803610500392160.

Albers, M. 2015. *Van Dale Middelgroot Woordenboek Nederlands*. 2nd ed. Utrecht: VBK Media.

Bialystok, E., G. Luk, K. F. Peets, and S. Yang. 2010. "Receptive Vocabulary Differences in Monolingual and Bilingual Children." *Bilingualism: Language and Cognition* 13 (4): 525–531. doi:10.1017/S1366728909990423.

Bosma, E., E. Blom, E. Hoekstra, and A. Versloot. 2016. "A Longitudinal Study on the Gradual Cognate Facilitation Effect in Bilingual Children's Frisian Receptive Vocabulary." *International Journal of Bilingual Education and Bilingualism* Advance Online Publication. doi:10.1080/13670050.2016.1254152.

Brenders, P., J. G. van Hell, and T. Dijkstra. 2011. "Word Recognition in Child Second Language Learners: Evidence from Cognates and False Friends." *Journal of Experimental Child Psychology* 109 (4): 383–396. doi:10.1016/J.JECP.2011.03.012.

Broersma, M. 2009. "Triggered Codeswitching between Cognate Languages." *Bilingualism: Language and Cognition* 12 (4): 447–462. doi:10.1017/S1366728909990204.

Brysbaert, M., and B. New. 2009. "Moving Beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English." *Behavior Research Methods* 41 (4): 977–990. doi:10.3758/BRM.41.4.977.

Buyl, A., and A. Housen. 2014. "Factors, Processes and Outcomes of Early Immersion Education in the Francophone Community in Belgium." *International Journal of Bilingual Education and Bilingualism* 17 (2): 178–196. doi:10.1080/13670050.2013.866627.

Cohen, C. 2016. "Relating Input Factors and Dual Language Proficiency in French-English Bilingual Children." *International Journal of Bilingual Education and Bilingualism* 19 (3): 296–313. doi:10.1080/13670050.2014.982506.

Crevecoeur, Y. C., M. D. Coyne, and D. B. McCoach. 2014. "English Language Learners and English-Only Learners' Response to Direct Vocabulary Instruction." *Reading & Writing Quarterly* 30 (1): 51–78. doi:10.1080/10573569.2013.758943.

Dahl, A., and M. Vulchanova. 2014. "Naturalistic Acquisition in an Early Language Classroom." *Frontiers in Psychology* 5: 329. doi:10.3389/fpsyg.2014.00329.

Dalton-Puffer, C. 2011. "Content-and-Language Integrated Learning: From Practice to Principles?" *Annual Review of Applied Linguistics* 31: 182–204. doi:10.1017/S0267190511000092.

de Graaff, R. 2015. "Vroeg of laat Engels in het basisonderwijs; Wat levert Het op?" *Levende Talen Tijdschrift* 16 (2): 3–15.

Dijkstra, T., K. Miwa, B. Brummelhuis, M. Sappelli, and H. Baayen. 2010. "How Cross-Language Similarity and Task Demands Affect Cognate Recognition." *Journal of Memory and Language* 62 (3): 284–301. doi:10.1016/J.JML.2009.12.003.

Dijkstra, T., and W. J. B. Van Heuven. 2002. "The Architecture of the Bilingual Word Recognition System: From Identification to Decision." *Bilingualism: Language and Cognition* 5 (3): 175–197. doi:10.1017/S1366728902003012.

Dongsun, Y., S. Yoon, and L. Jiyeon. 2016. "Receptive Vocabulary Assessment in Korean-English Bilingual Children: Cross-Linguistic Investigations." *Communication Sciences and Disorders* 21 (1): 131–138. doi:10.12963/csd.14186.

Dunn, L. M. 1959. *Peabody Picture Vocabulary Test*. Minneapolis: American Guidance Service.

Dunn, L. M., and D. M. Dunn. 2007. *PPVT-4 Manual*. Bloomington: NCS Pearson.

Enever, J., E. Krikhaar, E. Lindgren, L. Lopriore, G. Lundberg, J. Mihaljevic Djigunovic, C. Muñoz, M. Szpotowicz, and E. Tragant Mestres. 2011. *ELLiE: Early Language Learning in Europe*.

EP-Nuffic. (n.d.). Standaard tweetalig onderwijs Engels havo/vwo. https://www.nuffic.nl/bestanden/documenten/standaard-tweetalig-onderwijs-engels-havo-vwo.pdf

Gathercole, V. C. M., E. M. Thomas, and E. Hughes. 2008. "Designing a Normed Receptive Vocabulary Test for Bilingual Populations: A Model from Welsh." *International Journal of Bilingual Education and Bilingualism* 11 (6): 678–720. doi:10.1080/13670050802149283.

Goriot, C., M. Broersma, J. M. McQueen, S. Unsworth, and R. Van Hout. 2018. "Language Balance and Switching Ability in Children Acquiring English as a Second Language." *Journal of Experimental Child Psychology* 173: 168–186.

Heemskerk, J. S., and W. Zonneveld. 2000. *Uitspraakwoordenboek*. Houten: Het Spectrum.

Huang, B. H. 2016. "A Synthesis of Empirical Research on the Linguistic Outcomes of Early Foreign Language Instruction." *International Journal of Multilingualism* 13 (3): 257–273. doi:1080/14790718.2015.1066792.

Jenniskens, T., B. Leest, M. Wolbers, M. Bruggink, C. Dood, and E. Krikhaar. 2017. *Zicht op vroeg vreemdetalenonderwijs*. Nijmegen: KBA Nijmegen.

Jensen, S. 2017. "Gaming as an English Language Learning Resource among Young Children in Denmark." *Calico Journal* 34 (1): 1–19. doi:10.1558/cj.29519.

Jimenez Catalan, R. M., and M. Terrazas Gallego. 2005. "The Receptive Vocabulary of English Foreign Language Young Learners." *Journal of English Studies* 5–6: 173–191.

Keuleers, E., M. Brysbaert, and B. New. 2010. "SUBTLEX-NL: A New Measure for Dutch Word Frequency Based on Film Subtitles." *Behavior Research Methods* 42 (3): 643–650. doi:10.3758/BRM.42.3.643.

Kroll, J. F., and E. Stewart. 1994. "Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections between Bilingual Memory Representations." *Journal of Memory and Language* 33: 149–174.

Leśniewska, J., and F. Pichette. 2016. "Songs vs. Stories: Impact of Input Sources on ESL Vocabulary Acquisition by Preliterate Children." *International Journal of Bilingual Education and Bilingualism* 19 (1): 18–34. doi:10.1080/13670050.2014.960360.

Leśniewska, J., F. Pichette, and S. Béland. 2018. "First Language Test Bias? Comparing French-Speaking and Polish-Speaking Participants' Performance on the Peabody Picture Vocabulary Test." *Canadian Modern Language Review* 74 (1): 27–52. doi:10.3138/cmlr.3670.

Lindgren, E., and C. Muñoz. 2013. "The Influence of Exposure, Parents, and Linguistic Distance on Young European Learners' Foreign Language Comprehension." *International Journal of Multilingualism* 10 (1): 105–129. doi:10.1080/14790718.2012.679275.

Lobo, V. R. 2013. *Teaching L2 English at a Very Early Age: A Study of Dutch Schools*. Utrecht: LOT. http://www.lotschool.nl.

Longman. 2012. *Longman Dictionary of Contemporary English for Advanced Learners*. 6th ed. Harlow: Pearson Education.

McNamara, T. F. 2000. *Language Testing*. Oxford: Oxford University Press.
Merisuo-Storm, T. 2007. "Pupils' Attitudes Towards Foreign-Language Learning and the Development of Literacy Skills in Bilingual Education." *Teaching and Teacher Education* 23 (2): 226–235. doi:10.1016/j.tate.2006.04.024.
Moscoso del Prado Martín, F., R. Bertram, T. Häikiö, R. Schreuder, and R. Harald Baayen. 2004. "Morphological Family Size in a Morphologically Rich Language: The Case of Finnish Compared to Dutch and Hebrew." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30: 1271. doi:10.1037/0278-7393.30.6.1271.
Nikula, T. 2017. "CLIL: A European Approach to Bilingual Education." In *Second and Foreign Language Education. Encyclopedia of Language and Education*, edited by N. Van Deusen-Scholl, and S. May, 3rd ed., 111–124. Cham: Springer. doi:10.1007/978-3-319-02246-8_10.
Pérez, A. M., E. D. Peña, and L. M. Bedore. 2010. "Cognates Facilitate Word Recognition in Young Spanish-English Bilinguals' Test Performance." *Early Childhood Services* 4 (1): 55–67. doi:10.1044/1092-4388(2013/12-0056.
Poarch, G. J., and J. G. van Hell. 2012a. "Cross-Language Activation in Children's Speech Production: Evidence from Second Language Learners, Bilinguals, and Trilinguals." *Journal of Experimental Child Psychology* 111 (3): 419–438. doi:10.1016/J.JECP.2011.09.008.
Poarch, G. J., and J. G. van Hell. 2012b. "Executive Functions and Inhibitory Control in Multilingual Children: Evidence from Second-Language Learners, Bilinguals, and Trilinguals." *Journal of Experimental Child Psychology* 113 (4): 535–551. doi:10.1016/j.jecp.2012.06.013.
Potapova, I., H. K. Blumenfeld, and S. Pruitt-Lord. 2016. "Cognate Identification Methods: Impacts on the Cognate Advantage in Adult and Child Spanish-English Bilinguals." *International Journal of Bilingualism* 20 (6): 714–731. doi:10.1177/1367006915586586.
Schepens, J., T. Dijkstra, F. Grootjen, and W. J. B. van Heuven. 2013. "Cross-Language Distributions of High Frequency and Phonetically Similar Cognates." *PLoS ONE* 8 (5): e63006. doi:10.1371/journal.pone.0063006.
Schepens, J., F. van der Slik, and R. van Hout. 2016. "L1 and L2 Distance Effects in Learning L3 Dutch." *Language Learning* 66 (1): 224–256. doi:10.1111/lang.12150.
Steinlen, A., and T. Piske. 2013. "Academic Achievement of Children with and Without Migration Backgrounds in an Immersion Primary School: A Pilot Study." *Zeitschrift Für Anglistik Und Amerikanistik* 61 (3): 215–244. doi:10.1515/zaa-2013-0303.
Thijs, A., B. Trimbos, D. Tuin, M. Bodde, and R. de Graaff. 2011. *Engels in het Basisonderwijs*. Enschede: SLO.
Unsworth, S., L. Persson, T. Prins, and K. de Bot. 2015. "An Investigation of Factors Affecting Early Foreign Language Learning in the Netherlands." *Applied Linguistics* 36: 527–548. doi:10.1093/applin/amt052.
van der Leij, A., J. Bekebrede, and M. Kotterink. 2010. "Acquiring Reading and Vocabulary in Dutch and English: The Effect of Concurrent Instruction." *Reading and Writing* 23 (3): 415–434. doi:10.1007/s11145-009-9207-5.
van Heuven, W. J. B., P. Mandera, E. Keuleers, and M. Brysbaert. 2014. "SUBTLEX-UK: A New and Improved Word Frequency Database for British English." *Quarterly Journal of Experimental Psychology* 67 (6): 1176–1190. doi:10.1080/17470218.2013.850521.
Wells, J. C. 2008. *Longman Pronunciation Dictionary*. 3rd ed. Harlow: Pearson Education.
Wood, C., and V. Pena. 2015. "Lexical Considerations for Standardized Vocabulary Testing with Young Spanish-English Speakers." *Contemporary Issues in Communication Science and Disorders* 42: 202–215.