Article

LANGUAGE TESTING

Language Testing
1–22

DOI: 10.1177/0265532220961541
journals.sagepub.com/home/ltj


Assessing children's incremental word knowledge in the upper primary grades

Iris Monster
Radboud University, Netherlands

Agnes Tellings
Radboud University, Netherlands

William J. Burk
Radboud University, Netherlands

Jos Keuning
Cito Institute for Educational Measurement, Netherlands

Eliane Segers
Radboud University and the University of Twente, Netherlands

Ludo Verhoeven
Radboud University and Royal Dutch Kentalis, Netherlands

Abstract

Word knowledge acquisition is an incremental process that relies on exposure. As a result, word knowledge can broadly range from recognizing the word's lexical status, to knowing its meaning in context, and to knowing its meaning independent of context. The present study aimed to model incremental word knowledge in 1454 upper primary school children from grades 3 to 5 by investigating their abilities on three word knowledge tasks originating from the same set of 300 words: lexical decision, context decision, and definitional decision. A mixed-effects model showed significant differences in performance between tasks and between grades, and a significant interaction indicating that task differences were different for children in grade 5 compared to children in grades 3 and 4. In order to examine further the different task relation patterns at the word level, a cluster analysis was performed using the observed item means, which were corrected for the guessing chance. The analysis showed that for most words, recognition of its lexical status was easier than knowing its meaning in context, which in turn was easier than

Corresponding author:
Iris Monster, Behavioural Science Institute, Radboud University, Montessorilaan 3, P.O. Box 9104, Nijmegen, 6500 HE, Netherlands.
Email: i.monster@pwo.ru.nl

knowing its meaning independent of context. It is concluded that task relation patterns differ based on mean log frequency as a proxy of word exposure.

Keywords
Assessment, cluster analysis, incremental word knowledge, task relation patterns, upper primary grades

Word knowledge, defined as the size and quality of word meanings and word forms in the mental lexicon, is a crucial predictor of reading comprehension (e.g., Protopapas et al., 2007; Verhoeven et al., 2011). The accurate and fast retrieval of word meanings unburdens working memory and allows the brain to connect words into sentences, and sentences into larger passages of text (Perfetti & Stafura, 2014). According to the Lexical Quality Hypothesis (LQH), mental representations of words consist of orthographic, phonological, and semantic components that become incrementally connected in response to exposure to words in context (Perfetti, 2017). Since word knowledge increases through exposure, word knowledge can be partial. Researchers have proposed different models in which different aspects of word knowledge are described and are placed in an ordered sequence (see Christ, 2011). Such models of aspects of word knowledge hold for acquiring a second language (L2) as well (see, e.g., González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004). However, there are important differences between acquiring a first language (L1) vocabulary and an L2 vocabulary. For instance, Cremer et al. (2010) argued that adult L2 learners most likely resort to the semantic or conceptual knowledge developed in their first language (L1) in order to do vocabulary tasks in an L2, and that this holds also for children growing up bilingually, as long as they are familiar with the stimulus words in both the L1 and the L2. Typically, researchers studying word knowledge have not exhaustively examined the interrelations between different aspects of word knowledge, especially not with children who have just become literate and are still developing their first language.

An important question in vocabulary acquisition in the first language is how orthographic word knowledge is constructed as a consequence of literacy. As children learn to read, word decoding can be considered a self-teaching device that can help them decode words they have never seen or even heard before (Share, 2004). Upon repeated exposure to a novel word in context, increasing amounts of information are accumulated. Initially, there may be a vague notion of meaning or some recollection about the context in which the word was encountered until one is able to define the word correctly (Stahl, 2003). Accordingly, word knowledge can range from recognizing the word's lexical status, to knowing its meaning in context, and subsequently to knowing its meaning independent of context. In the present study, we assessed incremental L1 word knowledge in children in the upper primary grades (grades 3 to 5), examining differences at grade, task, and word levels.

Incremental word learning

According to the episodic model of word learning proposed by Reichle and Perfetti (2003), knowledge of phonological, orthographic, and semantic information (i.e., lexical

quality) increases as a function of encounters with words being used in different written and spoken contexts. Mental representations with high lexical quality contain precise and redundant information about the written and spoken form and the meaning of words. A flexible meaning representation is more generalized and contains a broader range of meaning dimensions to discriminate between semantically related words (Perfetti, 2017). Many studies corroborate the view that word knowledge grows incrementally through exposure (e.g., Frishkoff et al., 2011; Stahl, 2003). Beck et al. (2013) emphasized that the success in incremental word learning is influenced by the person's ability to infer meaning from context, on the one hand, and the degree of meaningful information that the context provides on the other hand.

Incremental word learning is also fostered by children's growing ability to generate and retrieve word definitions. In a study by Marinellie and Johnson (2004), it was found that children in grade 5 created more high-level definitions compared to children in grade 3. The older children more often than the younger children used an Aristotelian (i.e., dictionary) form of definition – such as "*coat* is a type of clothing you wear in the cold" – and they used a greater variety of word meaning features. In a subsequent study, Marinellie (2010) showed that older children (grade 4 and 5) were better at integrating semantic and syntactic properties from a newly learned definition into selecting the correct context sentence than younger children (grade 3). Thus, evidence indicates that word knowledge is accumulated over time and per exposure.

The degree of knowledge of a word (both in L1 and L2) is often described as the depth or quality of word knowledge (as opposed to breadth of knowledge; see the review by Schmitt, 2014). Depth of word knowledge can be conceptualized by breaking word knowledge down into sub-components. Nation (2001) distinguished three main components of L2 word knowledge that also apply to L1 word knowledge: knowledge of form (orthography and phonology), meaning (semantics), and use (different facets of pragmatics). Within each of these components, different subcategories of receptive knowledge and productive knowledge were discerned. In a classical study on accumulating knowledge of a word over time, Dale (1965) discriminated between four stages of word knowledge as a basis of assessing written vocabulary knowledge in school: (1) not knowing the word at all; (2) having heard the word before but not knowing the meaning; (3) being able to recognize the word in context and knowing some semantic features (e.g., knowing that *husting*s has something to do with elections); and (4) knowing the word well and remembering it. Furthermore, in a review study by Christ (2011) it was concluded that at least five consecutive levels of word knowledge can be distinguished, starting from no or incorrect knowledge, followed by partial or schematically related knowledge, contextual understanding, decontextual understanding, and paired understanding of contextual and decontextual knowledge.

Assessing incremental word knowledge

Based on the proposition of incremental aspects of word knowledge as described above, differential tasks have been proposed to measure word knowledge. They range from easy-to-explain tasks that measure relatively superficial knowledge of rather large sets of words in a relatively short amount of time, such as lexical decision tasks, to tasks that assess

deeper word knowledge in more complex tasks with smaller sets of words. Examples of the latter are synonym or antonym tasks, or tasks that require making or recognizing definitions of the target words (for an overview see Schoonen & Verhallen, 2008). According to Read (2000), three dimensions of vocabulary assessment can be distinguished: discreteness, selectivity, and context-dependency. He pointed out that a large proportion of vocabulary tests created are relatively discrete (i.e., they measure vocabulary as an independent construct versus embedded in some larger construct), use a representative set of selected words, and are context independent. For the selection of a representative set of target words, usually word frequency values would be consulted, as word frequency is a major predictor of word knowledge (e.g., Ibrahim et al., 2017; Moers et al., 2017; Verhoeven et al., 2011; Vermeer, 2001). Apart from word frequency and, amongst others, the age at which a word is acquired, the concreteness of the referent of the word, the morphological structure of the word, and the orthographic and phonological neighborhood size (i.e., how many words differ in only one grapheme or phoneme from that word) have also been identified as relevant predictors of word knowledge (see, e.g., Brysbaert, 2017; de Groot & Keijzer, 2000; Tellings et al., 2013; Yap & Balota, 2009).

Measuring word knowledge is complicated even further as studies have shown that subject characteristics, word properties, and task characteristics may interact. Coppens et al. (2013) found that word properties predicted word knowledge differently for hearing children and children with hearing loss in grades 3 through 6 and that the way they differed varied across different tasks. Diana and Reder (2006) found an interaction between task and word frequency in an undivided attention word recognition task versus a divided attention word recognition task, with university students. In the undivided attention condition, but not in the divided attention condition, they found a frequency effect such that participants more often incorrectly recognized words as having seen them in the study phase (i.e., more false alarms) when the words were more frequent. In another study with university students, word frequency and context variability (i.e., the number of different contexts in which a word is used) had different effects in two memorability measures (Aue et al., 2018). In an item recognition test, low-frequency items were recognized more often, whereas in an association recognition test high-frequency items were recognized more often. In both the item and association recognition tasks, items with low context variability were recognized more often.

Since word knowledge is a multi-faceted ability with many factors interacting with performance on different tasks, it may be challenging to develop a unidimensional scale of word knowledge. In the field of L2 word knowledge assessment, a study by Laufer and Goldstein (2004) tested words from different frequency ranges in four aspects: passive recognition, active recognition, passive recall, and active recall. They found a consistent hierarchy of these four aspects, but this may depend on the types of word knowledge being measured. Schmitt (2010) lists some of the issues that may arise with measuring word knowledge on a developmental scale, namely, variation in the order of aspects of word knowledge for different lexical items and different people, unclarity about the number of stages of word knowledge that should be distinguished, and vagueness about what should be considered a valid starting point and end point.

In summary, word learning, both in the L1 and in the L2, can be considered to be incremental, representing a variety in the quality of lexical representations between and

within children. Furthermore, it can be assumed that lexical quality may vary across words, depending on word properties. In assessing word knowledge, different tasks with different words thus tap into different aspects of lexical quality and depth of word knowledge. In the present study, the relationships between different aspects of word knowledge of young children who have just become literate are studied.

The present study

In the present study, an attempt was made to assess the incremental word knowledge of children in the upper grades of Dutch regular primary schools. Since the interrelations between aspects of word knowledge are not yet explored, we formulated three different assumptions in regard to these relations. The first assumption holds that word knowledge is a monolithic skill, meaning that children first know a few words (completely) and then increasingly learn more and more words (completely). The second assumption states that word knowledge builds up through incrementally acquiring different aspects of word knowledge, the acquisition of each aspect being conditional for the acquisition of the next one. The third assumption also postulates incremental growth of word knowledge, yet it does not assume a fixed order for all words. It rather assumes that aspects of word knowledge are interconnected in different ways for individual words or become differently interconnected at different phases of language development. The aim of the present study was to investigate which of these three assumptions best describes the relations between different aspects of word knowledge.

In order to assess children's incremental word knowledge, we distinguished three aspects of word knowledge based on the similarities between models of ordered aspects of word knowledge (Christ, 2011; Dale, 1965; Wesche & Paribakht, 1996). These three aspects were operationalized as three tasks used to assess word knowledge of 300 words, selected based on frequency in BasiLex (a 11.5 million word corpus of Dutch texts written for children, Tellings et al., 2014), in grades 3, 4, and 5 of elementary school.

The first task is a lexical decision task, which measures the first aspect of word knowledge described by most researchers (Christ, 2011; Dale, 1965; Wesche & Paribakht, 1996), namely, knowing that the letter string is a word in the target language. This context-independent task measures word recognition, but previous studies have shown that a lexical decision task evokes semantic information next to orthographic and/or phonological information (Coppens et al., 2011; Marcolini et al., 2009). The second task is a context decision task. Context decision refers to Christ's second level and Dale's third step, namely, understanding the meaning of a word in context. In this context-dependent task, the child has to recognize the usage of a word in a semantically correct linguistic context. The third and final task is a definitional decision task. In this context-independent task the child has to understand the correct definition of a word. Definitional cognition refers to Christ's third level, Dale's fourth step, and the notion of Perfetti (2017) that higher lexical quality includes a meaning representation that is decontextualized.

Two research questions were addressed.

1. *How does primary school children's performance on a lexical decision, context decision, and definitional decision task differ across grade levels?*

Given differences in task complexity, we expected that children would score higher on lexical decision than on context decision and higher on context decision than on definitional decision. We also expected that children in higher grades would score higher on each decision task than children in lower grades.

2. *How are lexical decision, context decision, and definitional decision related at the word level?*

When grouping target words together based on task relation patterns, we expected to find a general pattern of better performance on lexical decision than on context decision and on context decision than on definitional decision for most of the 300 words. However, given previous studies that show task and item property interactions (e.g., Coppens et al., 2013), we expected this pattern not to hold exclusively for all words. This expectation would provide support for our third assumption about the relations between different aspects of word knowledge.

Method

Participants

In total 1454 children in grades 3–5 of 23 Dutch elementary schools participated in this study: 456 children from grade 3; 478 children from grade 4; and 520 children from grade 5. Children were between 8 and 13 years old ($M = 10.1$ years, $SD = 11.8$ months, 52.4% girls); the majority of children were aged 8–11; and our sample included 27 children aged 12 and two children aged 13. Information about the native language or other acquired languages of the child was not collected, and could therefore not be controlled for within this study. In accordance with the APA ethical guidelines and the guidelines of our institute at the time of data collection (during spring 2017) the parents or guardians from all participating children were asked for their passive consent and none of them objected. All children had normal or corrected-to-normal vision and all schools were regular elementary schools (i.e., no special education schools).

Materials

Target words. The 300 target words were selected from the BasiLex 20,000 lemma list that is based on the BasiLex corpus (Tellings et al., 2014). BasiLex is a Dutch written-text child-input corpus of more than 11.5 million words that includes various types of texts written for children in different grades of primary school (e.g., textbooks, child literature, assessment tests, comics, subtitles for children's television shows, and websites). Words were drawn at random and then evaluated based on the part of speech and frequency. The part-of-speech criterion was included to ensure that the ratio of nouns, verbs, and adjectives was roughly equal to the ratio of these words in the BasiLex 20,000 lemma list (i.e., 62% nouns, 25% verbs, and 13% adjectives; in the BasiLex 20,000 list the distribution is 67%, 20%, and 14%, respectively). The frequency criterion checked

that at least 20% of the drawn words occurred at least four times in each of the grade sub-corpora 2 to 6. Frequencies of the 300 words ranged from 9 to 4828. The log transformed frequencies ranged from 0.95 to 3.68 ($M = 2.03$, $SD = 0.64$).

Lexical decision task. The lexical decision task comprised 300 pseudowords and 300 target words. Pseudowords were created by selecting for each target word another word from the BasiLex corpus that resembled that target word based on frequency, word class, and length. Then, one or two letters of the word were altered in order to create a phonologically and orthographically correct Dutch pseudoword.

When the lexical decision task started the program entered full-screen mode and children were presented with a black screen. Each iteration started with a fixation cross that was presented 750 ms and then after a delay of 750 ms the letter string (target word or pseudoword) would appear on screen for 5000 ms. Children answered by either pressing the "A" or the "L" button on a computer keyboard to indicate that the string of letters was a pseudoword or an existing Dutch word, respectively. If the child did not press any key or another key than the "A" or "L" keys, the response would not be recorded as false but as a non-response.

Context decision task. Based on previous studies (e.g., Coppens et al., 2011; Tellings et al., 2013), we developed a context decision task. A target word was used in four short sentences or phrases of which only one was semantically correct. On the screen children saw the target word and the four options with clickable bullets in front. This task was self-paced and children could only proceed to the next item once they had selected an option. A translated example, for target word *zijde* (Eng: silk), is as follows: "That dress is made from silk"; "That chicken has feet made from silk"; "I am eating at a table made from silk"; and "My pen is made from silk".

Definitional decision task. The definitional decision task was a self-paced word definition task. Definition tasks are often used to measure word knowledge (e.g., Ouellette, 2006; Vermeer, 2001). In our definition task a child was presented with a target word, a correct definition, and three alternatives. This format resembles the vocabulary part of the Shipley scale (Shipley et al., 2009) and part of the Dutch Cito Vocabulary Tests that most of the Dutch primary schools use as a standardized test for vocabulary. For the target word *zijde* (Eng: silk), the item was "A kind of fabric", "A building material", "A clap", and "A plane". All three tasks were created and presented to the children with PsyToolkit (Stoet, 2010, 2017).

Procedure

To reduce the burden for schools and children, the list of target words (and corresponding items on the three tasks) was randomly divided into three equally sized subsets of 100 target words each. These subsets represent three versions of the experiment and each child was presented with one version. Each subset was divided in two parts of 50 target words, as the experiment was divided over two sessions. Which half of the target words the children saw during the first or second session was counterbalanced across children that received the same version. Children were assigned a version and order by numbering

the list of names of participating children. The remainder of that number divided by six indicated which version and order the child received.

The sessions started with the lexical decision task and then the same target words were presented in the contextual decision task followed by the same target words in the definitional decision task. This order was decided upon as it was important that the lexical decision task came first; otherwise, children could learn from the other two tasks that a presented letter string was a word in the lexical decision task. As we expected that the definitional decision task would require deeper semantic knowledge than the contextual decision task and we wished to minimize learning effects, the contextual decision task was presented before the definitional decision task. If learning effects did occur, we assumed they were the same for all children.

At the start of the first session, children received a short oral presentation explaining all three tasks, including some examples. For the context decision task and the definitional decision task, children were instructed to select the answer that seemed "the best" to them. At the beginning of each task, the children received short, written (digital) instructions and they were given practice items including feedback based on their response. In the lexical decision task, there were four practice items (two words and two pseudowords) and for the context and definitional decision task children received two practice items. Children received new practice items in the second session. During the remainder of the task children did not receive any feedback. After the second task, children were given a short five-minute break and could do a connect-the-dots puzzle. During the tasks and after the full session, children received only verbal praise. On average, it took the children approximately one hour to complete a session, including a five-minute break.

Imageability

BasiLex log frequency and imageability estimates were used to describe the words in each cluster after performing a cluster analysis (see below). Using a digital task (Excel sheet) 20 adults (16 female, four male) were asked to rate the imageability of the 300 target words on a 7-point Likert-scale where 1 would represent low imageability and 7 high imageability. Age of the raters ranged between 20 and 36 years old and they all followed or finished a bachelor's programme. One male participant was removed from the results since he did not execute the task appropriately. The interrater consistency was computed using Cronbach's alpha ($\alpha = .96$ across 19 raters), which can be considered very high. For each word the imageability estimates were averaged across the 19 raters.

Plan of analysis

All statistics were computed in R, version 3.5.3 (R Core Team, 2019). The first research question regarding grade-level differences on the three lexical tasks was addressed with a generalized linear mixed-effects model (GLMM), performed using the glmer function in the R package lme4 (version 1.1.21; Bates et al., 2015). The responses to task items (correct/incorrect) were modelled as the outcome variable in this analysis. In the random effects structure of the generalized mixed-effects model, individual differences between

children were modelled as a random intercept. Individual differences in the associations between target words across the three tasks were modelled as random slopes. Task, grade level, and the interaction between these factors were included as fixed effects. The contrasts for both factors (task and grade level) were specified as dummy codes, with grade 4 and the contextual decision task used as reference categories. In addition, estimated marginal means were calculated and post-hoc comparisons (with Tukey adjustment) of the interaction effect were performed using the emmeans package (version 1.3.4; Lenth et al., 2019) to examine differences between tasks and grade levels.

The second research question involving the identification of homogeneous subgroups of words was addressed with a cluster analysis performed with the Mclust function of the mclust package in R, which is based on Gaussian finite mixture modelling (version 5.4.3; Scrucca et al., 2016). This analysis utilized the observed means of the items (the proportion of children that answered the item correctly) in order to identify subgroups of words that demonstrate different task relation patterns at the word level. By default, the Mclust function computes the fit, defined by the Bayesian Information Criterion (BIC), of 14 different model types (with varying covariance structures) with 1–9 clusters, which makes it possible to compare the statistical fit of 126 models. The implementation of the BIC in the Mclust package contains a negative component and should therefore be maximized (Fraley & Raftery, 1999; Schwarz, 1978). Since there were no a priori hypotheses about the covariance structure or the number of clusters, the model with the best statistical fit is described. The model with the best statistical fit maximizes the within-cluster homogeneity and the between-cluster heterogeneity based on the observed means of the words. By inspecting the mean observed difficulty per task and cluster, we can find out if the task difficulty order is the same for all words.

In order to provide some information about the average word in each cluster mean differences in log frequency and imageability between the clusters are examined with two one-way ANOVAs. Log frequency was used as it reflects exposure and imageability was used as it is a word characteristic that is not related to log frequency, $r(300) = .06$, $p = .286$. Imageability is fairly similar across languages.

Results

Descriptives

In Table 1, the means and standard deviations of the performance of the children on the three tasks are presented separately for each grade level. These task scores represent the percentage correct per task.

Table 2 presents the correlations between the three tasks based on the task scores of the children (below the diagonal) and based on the observed difficulty of the items of the words (above the diagonal), as well as the internal consistency of task items (on the diagonal, i.e., Cronbach's alpha). The correlations between the task scores of the children (presented below the diagonal) were computed using the percentage of items answered correctly per task for each child. The significant and positive correlations between tasks indicate that, on average, children with high scores on one task were also likely to have a relatively high score on another task. The correlations between the tasks at the word

Table 1. Means and standard deviations of the percentage of correct items obtained on the three tasks by children in the third, fourth, and fifth grade.

	Grade 3	Grade 4	Grade 5
n	456	478	520
LD (*SD*)	78.54 (9.57)	82.97 (8.49)	87.15 (7.12)
CD (*SD*)	70.65 (15.01)	76.41 (15.02)	82.58 (12.81)
DD (*SD*)	63.58 (14.68)	69.61 (15.17)	76.08 (12.98)

Note: LD = Lexical decision; CD = Contextual decision; DD = Definitional decision.

Table 2. Internal reliability and bivariate correlations of the three word knowledge tasks at the child and item level.

	LD	CD	DD
LD	.87	.66*	.48*
CD	.66*	.94	.64*
DD	.65*	.90*	.92

Note: Correlations among children's scores are presented below the diagonal ($n = 1454$); correlations among task items (averaged across versions) are presented above the diagonal ($n = 300$); Cronbach's alpha is presented on the diagonal per task.
LD = Lexical decision; CD = Contextual decision; DD = Definitional decision. * $p < .001$.

level (above the diagonal) were computed using the percentage of children that answered the item of the word correctly per task. The significant and positive correlations at the word level indicate that, on average, items that were correctly answered by children on one task were also likely to be correctly answered by the same children the other tasks.

In order to provide additional information about the tasks at item level a Rasch model was estimated using software from Verhelst et al. (1995). This analysis yielded a metric on which both item difficulty and child ability are located (on a logit scale). The fit of the model was evaluated by comparing the observed item difficulty to the model-predicted item difficulty and the observed total sum scores to the model-predicted total sum scores. The correlation between observed and predicted item difficulty was very strong and significant, $r(898) = .998$, $p < .001$. The correlation between the observed and predicted total scores was also very strong and significant, $r(1452) = .980$, $p < .001$. The predictions by the Rasch model could thus be considered accurate. Therefore, in the next step, ggplot2 (Wickham, 2016) version 3.2.1 was used to create a so-called Wright person–item map (see Figure 1). On the *y*-axis the child (left-hand side) and item (right-hand side) parameters are shown. A negative value indicates a lower child ability or a lower item difficulty and a positive value indicates a higher child ability or a higher item difficulty. As can be seen, the range of the item difficulties is large as it surpasses the range of the child abilities. Furthermore, as most of the child abilities are larger than 0.0 and a large portion of the item difficulties below 0.0, we may conclude that the tasks were relatively easy for many children.

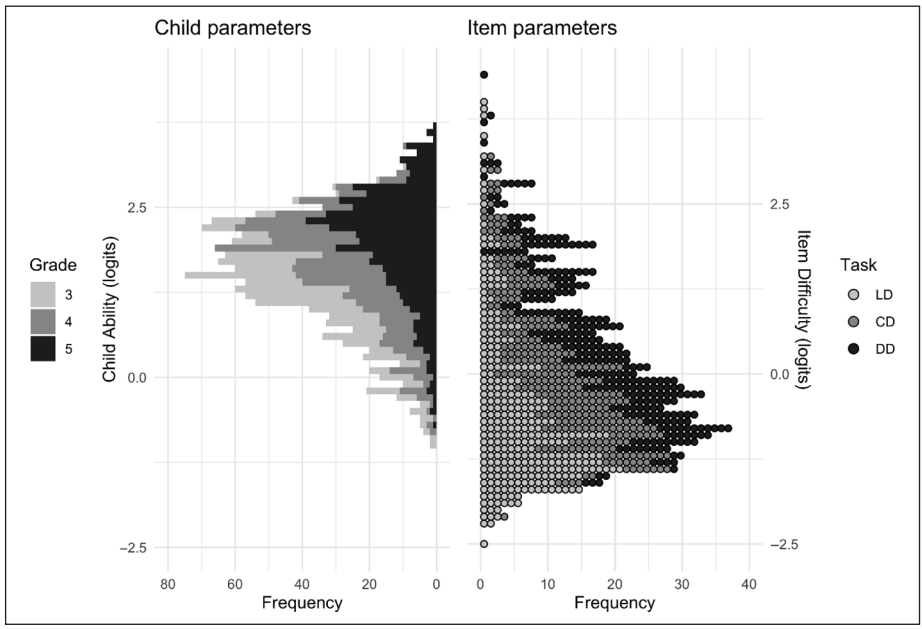

Figure 1. Wright map of child ability estimates and item difficulty estimated on the same scale.

Research question 1: Differences across tasks and grade levels

A generalized linear mixed-effects model was conducted to investigate the effects of grade level and task type on word knowledge. The log odds ratios and 95% confidence intervals for the model estimates are presented in Table 3. In Table 3 the random effects, variances, and intra-class correlation coefficients of participant and word are also presented.

The contrasts involving the main effects for task type (CD vs. LD and CD vs. DD) and grade level (grade 4 vs. grade 3 and grade 4 vs. grade 5) emerged as statistically significant, but these differences were qualified by an interaction between these factors. A multiple comparisons matrix was computed to examine the differences between grade levels across all three tasks (the interaction effect). The results of these comparisons are presented in Table 4. The predicted probabilities and 95% confidence intervals of performance were plotted separately for each grade and task level to visualize the mean-level differences (see Figure 2).

The results of the multiple comparisons matrix of the interaction effect indicate that for the children in grade 5, the pattern of mean-level differences was different compared to the pattern of mean-level differences for the children in grades 3 and 4. The positive coefficients of the LD – CD contrasts with grade contrasts 3 – 5 and 4 – 5 indicate that for the children in grades 3 and 4, compared to the children in grade 5, the lexical decision task mean probability to select the correct answer was relatively higher than the context decision task mean probability to select the correct answer. So, the difference in

Table 3. Log odds ratios and 95% confidence intervals for task type and grade level as predictors of performance.

	Performance on decision task		
Predictors	*Log-odds ratio*	*95% CIs*	*p*
Intercept	1.55	[1.40, 1.70]	**<.001**
LD	0.52	[0.41, 0.64]	**<.001**
DD	−0.44	[−0.56, −0.33]	**<.001**
Grade 3	−0.41	[−0.51, −0.30]	**<.001**
Grade 5	0.51	[0.40, 0.61]	**<.001**
LD × Grade 3	−0.00	[−0.05, 0.05]	.903
DD × Grade 3	0.03	[−0.02, 0.07]	.269
LD × Grade 5	−0.06	[−0.12, −0.01]	**.023**
DD × Grade 5	−0.06	[−0.11, −0.01]	**.019**
Random effects			
σ^2	3.29		
$\tau_{00\ ppn}$	0.59		
$\tau_{00\ word}$	1.34		
$\tau_{11\ word.task.CD}$	1.29		
$\tau_{11\ word.task.DD}$	1.41		
$\rho_{01\ word.task.CD}$	0.66		
$\rho_{01\ word.task.DD}$	0.49		
ICC_{ppn}	0.11		
ICC_{word}	0.26		
Observations	432158		
Marginal R^2 / Conditional R^2	0.055 / 0.353		

Note: LD = Lexical decision; CD = Context decision; DD = Definitional decision; Reference is grade 4 and task CD (Context decision); σ^2 indicates the within-group (residual) variance; τ_{00} indicates the variation between individual intercepts and the average intercept for children (ppn) and words (word); τ_{11} indicates the variation between individual task slopes and the average slope; ρ_{01} indicates the random-intercept-slope-correlation.

odds ratio between the LD and CD is greater for grade 3 compared to grade 5 and for grade 4 compared to grade 5. The negative coefficients of the CD – DD contrasts with grade contrasts 3 – 5 and 4 – 5 indicate that for the childre

n in grade 3 and 4, compared to the children in grade 5, the context decision task mean probability to select the correct answer was relatively lower than the definitional decision task mean probability to select the correct answer. So, the difference in odds ratio between the CD and DD is smaller for grade 3 compared to grade 5 and for grade 4 compared to grade 5.

Research question 2: Task relation patterns

Cluster analyses were performed with the observed means per task and for each word (the ratio of children that answered an item correctly, from here on referred to as the task

Table 4. Multiple comparisons matrix for the interaction effect (differences between tasks across different grades).

Task	Grade	Log odds ratio	SE	*z* ratio	*p*
LD – CD	3 – 4	−0.003	0.026	−0.122	.903
LD – CD	3 – 5	0.060	0.027	2.203	**.028**
LD – CD	4 – 5	0.063	0.028	2.271	**.023**
LD – DD	3 – 4	−0.030	0.026	−1.164	.244
LD – DD	3 – 5	−0.025	0.026	−0.962	.336
LD – DD	4 – 5	0.004	0.027	0.164	.870
CD – DD	3 – 4	−0.027	0.024	−1.106	.269
CD – DD	3 – 5	−0.086	0.025	−3.463	**.001**
CD – DD	4 – 5	−0.059	0.025	−2.355	**.019**

Note: LD = Lexical decision; CD = Context decision; DD = Definitional decision.

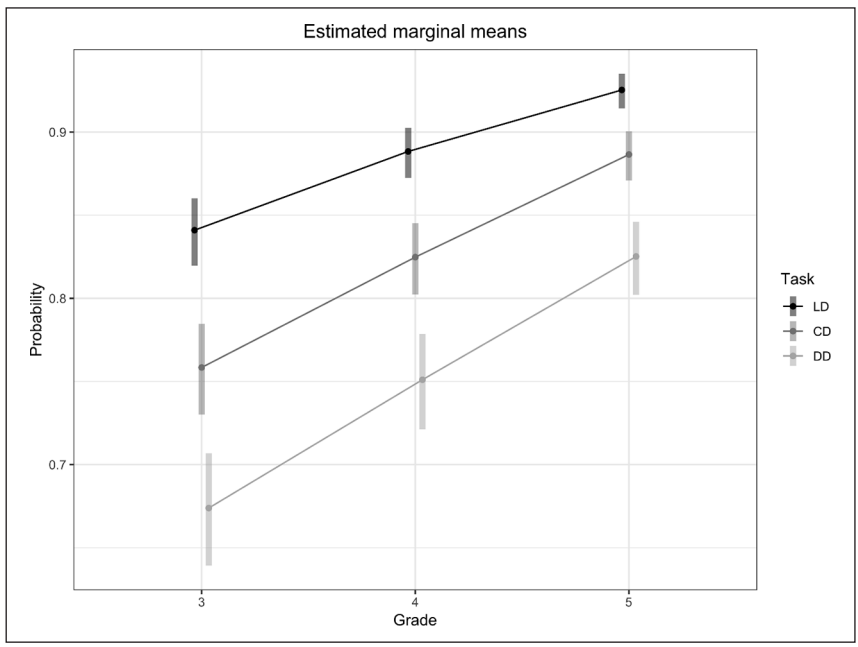

Figure 2. Estimated marginal means and 95% confidence intervals of the probability of accuracy as a function of grade level and task.

score). The model with the highest BIC was a five-cluster model with a covariance structure that had a diagonal distribution and variable volume and shape (BIC = 1332.917). Upon inspection of the cluster means of the task scores, we noticed that the difference in the guessing chance between the lexical decision task (two options) and the context and definitional decision task (four options) distorted the pattern. Namely,

Table 5. Means of the corrected task scores of the five-cluster model and means, standard deviation, minimum and maximum of classification certainty.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Overall *M (SD)*
n	49	109	32	42	68	
Mean-corrected task scores						
LD	0.092 [a]	0.832 [c]	0.623 [b]	0.887 [c]	0.717 [b]	0.662 (0.332)
CD	0.334 [a]	0.815 [c]	0.813 [c]	0.904 [c]	0.577 [b]	0.691 (0.236)
DD	0.263 [a]	0.698 [b]	0.791 [c]	0.872 [c]	0.440 [a]	0.601 (0.266)
Classification certainty						
M	0.923	0.842	0.884	0.892	0.851	0.869 (0.136)
SD	0.144	0.123	0.132	0.115	0.153	
Min	0.513	0.469	0.579	0.558	0.367	
Max	1.0	0.984	0.998	0.983	0.991	

Note: LD = Lexical decision; CD = Contextual decision; DD = Definitional decision.
Comparisons of mean-corrected task scores are facilitated by indicating means as either below average [a], average [b], or above average [c] based on the range overall task mean ± 0.5 * standard deviation.

in the first cluster ($n = 49$) the mean LD score was slightly above chance levels (.54 and the chance level was .5 for this two-option task) and the means for the other two tasks were well above the chance level (.50 and .45 for the CD and DD respectively; both tasks had four options so the chance level was .25). Therefore, the task scores were corrected for the guessing chance using a standard correction formula. For the LD items this was (C–(N/2)) / (0.5N) and for CD and DD items: (C–(N/4)) / (0.75N), where C is the number of children that responded correctly to the item and N is the number of children presented with the item. The correlation between original task scores and these corrected task scores per task is exactly 1.0.

The most parsimonious cluster solution with the corrected task scores was also a five-cluster model (BIC = 517.824) with the same covariance structure and word classifications as the best fitting model with the original task scores. Only for the first cluster did using corrected task scores result in a different task relation pattern (compared to the cluster 1 task relation pattern of the original task scores), where the LD mean of the corrected scores was considerably lower than the CD and DD means of the corrected scores.

In Table 5, the means of the corrected task scores per cluster are presented. Table 5 also contains the mean, standard deviation, minimum, and maximum per cluster of the word classification certainties (the probability of belonging to a cluster). Mean-corrected task scores are compared within and between clusters by computing the mean and standard deviation of the corrected task scores per task (the overall mean and standard deviation per task are presented in Table 5). Task cluster means that fall within the range of mean ± 0.5 * standard deviation (of the respective task) are considered to be average. Task cluster means outside of this range are considered to be below or above average.

Table 6. Mean-level differences on log frequency and imageability of the words in the three cluster solutions.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Log frequency					
M	1.48	2.19	1.80	2.66	1.90
SD	0.45	0.59	0.38	0.55	0.54
CLD	A	C	B	D	B
Imageability					
M	3.35	4.15	4.78	4.54	3.32
SD	1.33	1.39	1.03	1.48	1.54
CLD	A	B	B	B	A

Note: CLD (Compact Letter Display) indicates significant differences between cluster means. Within rows, clusters with different letters significantly differ from each other ($p < .01$).

The first cluster of the five-cluster model contained 49 words (16.3%) that, on average, have a below-average lexical decision-corrected score. The mean-corrected task scores of the context decision and definitional decision task were somewhat higher, but also below average. Cluster 2 contained 109 words (36.3%) with, on average, above-average corrected scores on the lexical decision and context decision task, the mean-corrected score of the definitional decision task was average. Cluster 3 contained 32 words (10.7%) with an average mean score on the lexical decision task and above average mean-corrected task scores on the context decision and definitional decision tasks. Cluster 4 contained 42 words (14%) with above-average mean-corrected task scores on all three tasks. The fifth cluster contained 68 words (22.7%) with average corrected task score means on the lexical decision and context decision tasks and a below-average definitional decision-corrected task score mean.

In order to explain further the differences between the task relation patterns of the clusters, two one-way ANOVAs (with Welch correction) were performed to test whether the words in the clusters differ on their frequency in the BasiLex corpus (Tellings et al., 2014) and on their imageability. For these analyses, the log frequencies were used, since frequency distributions of words in large text corpora typically follow Zipf distributions (frequency drops exponentially by rank, Baayen, 2002) and log-transforming the frequencies reduces the skewness. For post-hoc comparisons, Welch's dependent samples *t*-tests were performed with the Benjamini-Hochberg correction. Means and standard deviations of the log frequency and imageability estimates of words in each cluster are presented in Table 6. Significant differences between clusters are indicated in Table 6 using Compact Letter Display (CLD).

The Welch's ANOVA testing differences in mean log frequency between the clusters was statistically significant, $F(4, 119.21) = 36.348$, $p < .001$. The post-hoc comparisons indicated that all clusters except clusters 3 and 5 differed significantly from each other based on mean log frequency. Cluster 1 contained words with the lowest mean log frequency, followed by words in clusters 3 and 5 (that did not differ from each other). The mean log frequency of the words in cluster 2 was significantly higher than the mean log

frequency of clusters 3, 5, and 1. Last, the mean log frequency of the words in cluster 4 was significantly higher than the mean log frequencies of the other clusters.

The Welch's ANOVA testing differences in mean imageability between the clusters was also statistically significant, $F(4, 116.41) = 12.267$, $p < .001$. The post hoc comparisons showed that mean imageability significantly differed for clusters 1 and 5 versus clusters 2, 3, and 4. The mean imageabilities for the words in clusters 2, 3, and 4 did not differ significantly from each other, but were significantly higher than the mean imageabilities of clusters 1 and 5.

Discussion

The aim of this study was to investigate the incremental knowledge of the same set of words as measured by tasks measuring three aspects of word knowledge in children across the intermediate primary grades. Therefore, a study was conducted with Dutch children from grade 3, 4, and 5 who received 300 target words in three different tasks tapping incremental levels of word knowledge. These tasks were a lexical decision task tapping the recognition of a word's lexical status, a context decision task tapping the identification of the meaning of a word in a sentence context, and a definitional decision task tapping the identification of the decontextualized definition of a word. Based on previous studies, we expected to find significant performance differences based on task difficulty and grade level. A mixed-effects model confirmed our hypotheses; there were mean-level differences at both task and grade level, but these differences were qualified by a significant interaction effect. The significant interaction effect between grade level and task indicated that for children in grade 5 the pattern of mean-level task differences was different compared to the children in grades 3 and 4. These findings support the assumption that word knowledge acquisition of primary school children is an incremental process in which, in general, word knowledge grows from word recognition to contextualized meaning identification, and in the end decontextualized meaning identification (Reichle & Perfetti, 2003).

This study also explored the relations between the tasks at the word level using a cluster analysis to identify distinct subgroups of words based on the observed means of each item that were corrected for differences in the guessing chance. The most parsimonious cluster solution (maximizing within-cluster homogeneity and between-cluster heterogeneity) was found for a model with five clusters. The cluster means of the corrected task scores indicated that there are different task relation patterns among the set of 300 words on log frequency (in Basilex) and imageability.

In general, words with the highest mean log frequency and with high mean imageability estimates (cluster 4) were easy on all tasks with very little differences in task difficulty. Mean-corrected task scores of the lexical and context decision tasks were also above average for the words in cluster 2, whereas the definitional decision mean-corrected task score was average. The average log frequency of these words was also lower compared to the words in cluster 4. On average, words in cluster 5 showed average lexical and context decision mean-corrected task scores and a below-average definitional decision mean-corrected task score. Words in cluster 5 showed a significantly lower mean log frequency and lower mean imageability compared to the words in clusters 2

and 4. Cluster 4 could be interpreted as a ceiling-effect cluster with little to no difference in task complexity, but for the words in the other two clusters the general pattern held that lexical decision was easier than context decision, which was easier than definitional decision.

The other two clusters (3 and 1, with a combined size of 81 words) showed a different average task relation pattern. In both clusters, the mean of the lexical decision task scores was lower than the means of the other tasks. This indicated that children performed relatively better when presented with contextual cues, but still poorly compared to words in other clusters. The words in cluster 1 (with below-average mean-corrected task scores) also showed the lowest mean log frequency and were (together with the words in cluster 5) low in mean imageability. Remarkably, the words in cluster 3 showed an average mean-corrected task score on the lexical decision task and above-average mean-corrected task scores of the context decision task and definitional decision task. On average, the words in cluster 3 showed a fairly low frequency (ranked second lowest together with cluster 5), but a high imageability mean.

It is interesting to note that there were words for which lexical decision was relatively difficult compared to context decision and definitional decision. This finding can be explained from the basic idea that words are stored in a mental network as has been described by Meara and Wolter (2004). In this intra-connected network model, lexical items are connected to associated items. So presumably, children have not yet been exposed to some words often and have never seen the written form of the word yet. Since the connections between the mental representations from orthographic form to phonological form to (some) retrievable semantic information of the word have not been developed (enough), children struggle to recognize the word in isolation. However, when presented with semantically related words in a linguistic context and in a definitional decision task, the activation of these semantically related words in the intra-connected mental lexicon facilitates retrieving semantic knowledge of the target word. Furthermore, these findings support usage-based approaches of language acquisition (Langacker, 2000) that imply that linguistic context helps in solving language tasks. Usage-based approaches start from the assumption that children acquire both word knowledge and grammar based on multiple exposures to phrases such as “Drink your X” or “Please give me the long X”. Thus, words are always learned together with other words with which they frequently co-occur (Tomasello, 2000).

In line with the three possible assumptions about the dimensionality of word knowledge that were proposed in the introduction, it can be concluded that word recognition, contextualized word meaning, and decontextualized word meaning reflect incremental aspects of word knowledge. However, it was also shown that there are words in the low frequency domain for which accurately recognizing a word is not conditional in order to identify its correct use in context or its definition. Most probably, accessing the mental representation of these words is facilitated by the contextual cues in the context decision task and definitional decision task. The high dependency on word exposure and the compensatory function of context in the domain of low-frequency words are best aligned with our third assumption.

It is important to note that in this study the Dutch word knowledge of a random sample of children in Dutch primary school was assessed to foster generalizability of study

outcomes. However, it should be recognized that about 26.9% of children in Dutch primary school have a migration background (Central Bureau for Statistics, 2020) and have probably learnt Dutch as a second language, and that this information was not assessed in the current study. Previous research has shown that although bilingual children have (on average) lower vocabulary scores in their L2 compared to the scores of monolingual children, there is no evidence of differences between monolingual and bilingual children with regard to the order of acquisition of lexical items (van den Bosch et al., 2019; Vermeer, 2001). Based on this finding, it can be assumed that the grade and task level differences that we found, would also apply to the bilingual children. Furthermore, there is evidence that frequencies and imageability scores of lexical items across languages are highly comparable. Rofes et al. (2018) found correlations varying from $r = .31$ to $r = .92$, when comparing imageability ratings for 13 European languages. At the child level, individual differences may affect the performance of the child to a specific item. For example, it has been shown that cognate items with which the child is familiar in both languages may facilitate performance (see Tonzar et al., 2009). Therefore, future research on L2 acquisition and testing should take into account that linguistic overlap and differences may affect the prediction of word learning parameters in bilingual children (e.g., Benigno & de Jong, 2019).

Evidently, the present study can only be seen as a first step in uncovering the mechanisms involved in incremental word learning. Several limitations apply to our study. First of all, it should be noted that a cross-sectional design was followed in examining the role of grade-level effects on word learning. In order to be able to take a full developmental perspective on the stages of incremental word knowledge, longitudinal studies are needed in which different lexical tasks are provided to the same children over time. A second limitation is that the present study used multiple-choice tasks to measure children's lexical knowledge, which is a format that yields less information than, for instance, open question formats. However, it should be noted that multiple-choice items limit the information in context decision and definitional decision tasks and thus produce less noise. A third limitation is that the results from the cluster analyses should be interpreted cautiously as they reflect general task relation patterns, but cannot be interpreted as patterns that accurately describe each word in a cluster. The cluster analyses served as exploratory analyses to examine the heterogeneity in task relation patterns among words. The comparisons of mean log frequency and imageability were intended to provide background information about the average word in each cluster and not intended as an exhaustive analysis of the word characteristics that may cause differences between clusters.

This study also has practical implications. The results show that, in general, word recognition is easier than recognizing word meaning in context, which is easier than recognizing a word's definition. However, the relations between these three measures can differ between words with different frequencies. This finding has implications for both word knowledge testers and word knowledge trainers. For low-frequency words, it should not be assumed that children can decode them even if the children's general level of word decoding in itself is sufficient. Word knowledge testers should take into consideration that task type and word frequency interact, which may lead to different estimates of word difficulty depending on the task type. The relation between the context decision

task and definitional decision task indicated that, in general, generalizing to a context-independent meaning representation of a word is a stage beyond recognizing meaning in context. Therefore, exercises focused on the generation of semantic webs may facilitate word definition skills through the perspective of robust storage of word meanings in semantic memory.

In conclusion, the present study sheds light on the incremental building up of word knowledge in children in the upper primary grades. For most words, it was demonstrated that recognition of its lexical status was easier than knowing its meaning in context, which in turn was easier than knowing its meaning independent of context. In particular, the comparison of task difficulty at the word level showed that, for words with a low frequency, contextual cues are a valuable way to access a word in the mental lexicon and help in solving language tasks.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Iris Monster https://orcid.org/0000-0001-6892-2050

References

Aue, W. R., Fontaine, J. M., & Criss, A. H. (2018). Examining the role of context variability in memory for items and associations. *Memory & Cognition*, 1–15. https://doi.org/10.3758/s13421-018-0813-9

Baayen, R. H. (2002). *Word frequency distributions* (*Vol. 18*). Springer Science & Business Media. https://doi.org/10.1007/978-94-010-0844-0

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Benigno, V., & de Jong, J. (2019). Linking vocabulary to the CEFR and the Global Scale of English: A psychometric model. In A. Hutha, G. Erickson & N. Figueras (Eds.), *Developments in language education. A memorial volume in honour of Sauli Takala* (pp. 8–29). University Printing House.

Brysbaert, M. (2017). Age of acquisition ratings score better on criterion validity than frequency trajectory or ratings "corrected" for frequency. *The Quarterly Journal of Experimental Psychology*, *70*(7), 1129–1139. https://doi.org/10.1080/17470218.2016.1172097

Central Bureau for Statistics (2020, May). *Leerlingen in (speciaal) basisonderwijs; migratieachtergrond, woonregio* [Data file]. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83295NED/table?ts=1591612652210

Christ, T. (2011). Moving past "right" or "wrong" toward a continuum of young children's semantic knowledge. *Journal of Literacy Research*, *43*(2), 130–158. https://doi.org/10.1177/1086296X11403267

Coppens, K. M., Tellings, A., Verhoeven, L., & Schreuder, R. (2011). Depth of reading vocabulary in hearing and hearing-impaired children. *Reading and Writing*, *24*(4), 463–477. https://doi.org/10.1007/s11145-010-9237-z

Coppens, K. M., Tellings, A., Verhoeven, L., & Schreuder, R. (2013). Reading vocabulary in children with and without hearing loss: The roles of task and word type. *Journal of Speech, Language, and Hearing Research*, *56*, 654–666. https://doi.org/10.1044/1092-4388(2012/11-0138)

Cremer, M., Dingshoff, D., de Beer, M., & Schoonen, R. (2010). Do word associations assess word knowledge? A comparison of L1 and L2, child and adult word associations. *International Journal of Bilingualism*, *15*(2), 187–204.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, *42*(8), 895–948. https://www.jstor.org/stable/41385916

De Groot, A. M., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1–56. https://doi.org/10.1111/0023-8333.00110

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 805–815. https://doi.org/10.1037/0278-7393.32.4.805

Fraley, C., & Raftery, A. E. (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification*, *16*(2), 297–306. https://doi.org/10.1007/s003579900058

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2011). Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading*, *15*(1), 71–91. https://doi.org/10.1080/10888438.2011.539076

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Ibrahim, A., Cowell, P. E., & Varley, R. A. (2017). Word frequency predicts translation asymmetry. *Journal of Memory and Language*, *95*, 49–67. https://doi.org/10.1016/j.jml.2017.02.001

Langacker, R. W. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-based models of language* (pp 1–63). CSLI.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Emmeans: Estimated marginal means, aka least-squares means. https://CRAN.R-project.org/package=emmeans.

Marcolini, S., Burani, C., & Colombo, L. (2009). Lexical effects on children's pseudoword reading in a transparent orthography. *Reading and Writing*, *22*(5), 531–544. https://doi.org/10.1007/s11145-008-9123-0

Marinellie, S. A. (2010). The understanding of word definitions in school-age children. *Journal of Psycholinguistic Research*, *39*(3), 179–197. https://doi.org/10.1007/s10936-009-9132-4

Marinellie, S. A., & Johnson, C. J. (2004). Nouns and verbs: A comparison of definitional style. *Journal of Psycholinguistic Research*, *33*(3), 217–235. https://doi.org/10.1023/B:JOPR.0000027963.80639.88

Meara, P., & Wolter, B. (2004). V_Links: Beyond vocabulary depth. *Angles on the English Speaking World*, *4*, 85–96. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.716.3948&rep=rep1&type=pdf

Moers, C., Meyer, A., & Janse, E. (2017). Effects of word frequency and transitional probability on word reading durations of younger and older speakers. *Language and Speech*, *60*(2), 289–317. https://doi.org/10.1177/0023830916649215

Nation, I. S. (2001). *Learning vocabulary in another language*. Ernst Klett Sprachen. https://doi.org/10.1017/CBO9781139524759

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554–566. https://doi.org/10.1037/0022-0663.98.3.554

Perfetti, C. A. (2017). Lexical quality revisited. In E. Segers & P. van den Broek (Eds.), *Developmental perspectives in written language and literacy: In honor of Ludo Verhoeven* (pp. 51–67). John Benjamins. https://doi.org/10.1075/z.206.04per

Perfetti, C. A., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Protopapas, A., Sideridis, G. D., Mouzaki, A., & Simos, P. G. (2007). Development of lexical mediation in the relation between reading comprehension and word reading skills in Greek. *Scientific Studies of Reading*, *11*(3), 165–197. https://doi.org/10.1080/10888430701344322

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Read, J. (2000). *Assessing vocabulary* (pp. 1–16). Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Reichle, E. D., & Perfetti, C. A. (2003). Morphology in word identification: A word experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading, 7*, 219–238. https://doi.org/10.1207/S1532799XSSR0703_2

Rofes, A., Zakariás, L., Ceder, K., Lind, M., Johansson, M. B., De Aguiar, V., . . . Sacristán, C. H. (2018). Imageability ratings across languages. *Behavior Research Methods*, *50*(3), 1187–1197. https://doi.org/10.1207/S1532799XSSR0703_2

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer. https://doi.org/10.1207/S1532799XSSR0703_2

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*(2), 211–236. https://doi.org/10.1177/0265532207086782

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016) mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal 8*(1), 205–233. https://journal.r-project.org/archive/2016-1/scrucca-fop-murphy-etal.pdf

Share, D. L. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology*, *87*(4), 267–298. https://doi.org/10.1016/j.jecp.2004.01.001

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). *Shipley-2 Manual*. Western Psychological Services. https://doi.org/10.1037/t48948-000

Stahl, S. (2003). Words are learned incrementally over multiple exposures. *American Educator, 27*, 18–22.

Stoet, G. (2010). PsyToolkit – A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Tellings, A., Coppens, K., Gelissen, J., & Schreuder, R. (2013). Clusters of word properties as predictors of elementary school children's performance on two word tasks. *Applied Psycholinguistics*, *34*(3), 461–481. https://doi.org/10.1017/S014271641100083X

Tellings, A., Hulsbosch, M., Vermeer, A., & van den Bosch, A. (2014). BasiLex: An 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, *4*, 191–208. https://repository.ubn.ru.nl/bitstream/handle/2066/134845/134845.pdf?sequence=1

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1/2), 61–82. https://doi.org/10.1515/cogl.2001.012

Tonzar, C., Lotto, L., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning*, *59*(3), 623–646. https://doi.org/10.1111/j.1467-9922.2009.00519.x

van den Bosch, L. J., Segers, E., & Verhoeven, L. (2019). The role of linguistic diversity in the prediction of early reading comprehension: A quantile regression approach. *Scientific Studies of Reading*, *23*(3), 203–219. https://doi.org/10.1080/10888438.2018.1509864

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One Parametric Logistic Model OPLM*. CITO. https://doi.org/10.1007/978-1-4612-4230-7_12

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading*, *15*(1), 8–25. https://doi.org/10.1080/10888438.2011.536125

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*(2), 217–234. https://doi.org/10.1017/S0142716401002041

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13–40. https://doi.org/10.3138/cmlr.53.1.13

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://doi.org/10.1007/978-3-319-24277-4

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529. https://doi.org/10.1016/j.jml.2009.02.001