

Examining incidental word learning during reading in children:

The role of context

Abstract

From mid-childhood onwards, children learn hundreds of new words every year incidentally through reading. Yet little is known about this process, and the circumstances in which vocabulary acquisition is maximised. We examined whether encountering novel words in semantically diverse rather than semantically uniform contexts led to better learning. Children aged 10-11 read sentences containing novel words while their eye movements were monitored. Results showed a reduction in reading times over exposure for all children, but especially those with good reading comprehension. There was no difference in reading times or in offline post-test performance for words encountered in semantically diverse and uniform contexts but diversity did interact with reading comprehension skill. Contextual informativeness also affected reading behaviour. We conclude that children acquire word knowledge from incidental reading, that children with better comprehension skills are more efficient and competent learners, and that although varying the semantic diversity of the reading episodes did not improve learning per se in our laboratory manipulation of diversity, diversity did affect reading behaviour in less direct ways.

Key words

Word learning

Reading

Eye movements

Semantic diversity

Children

Word count = 10,475

Introduction

From mid-childhood onwards, children acquire hundreds of new words every year from their reading experience (Nagy & Anderson, 1984). Once the basics of learning to read are in place, children rapidly learn new orthographic forms, with evidence of some orthographic learning following a single exposure (Nation, Angell & Castles, 2007; Nation & Castles, in press; Share, 2004). Learning the meaning of a new word is more complex, with multiple exposures needed to develop a full understanding of its meaning. Although this kind of incidental word learning through reading is commonplace, we know relatively little about how the gradual acquisition of word knowledge via reading occurs, or the contextual factors that help or hinder the acquisition of new word meanings.

Nation (2017) argued that reading experience provides exposure to words in different contexts or episodes which over time sum to a rich database about their lexical history within an individual's experience. The product of these encounters is lexical quality, defined as variation in the extent to which the mental representation of a word specifies its meaning and form (Perfetti, 2007). In this paper we investigate the nature of contextual experience and ask whether word learning is enhanced when reading experience is varied rather than maintained.

Contextual diversity and lexical processing

Few studies have explored contextual experience and word learning directly. Relevant however is a growing literature on contextual diversity and lexical processing. Adelman, Brown and Quesada (2006) operationalized contextual diversity as the number of unique documents a word appears in across a large corpus. Document count

predicted lexical decision time in skilled readers, accounting for more unique variance than word frequency, demonstrating that it is not simply how often a word occurs in language that is critical, but the number of different contexts it appears in. Document count also predicts lexical decision performance in children (Perea, Soares & Comesaña, 2013) and the general effect has been replicated using eye movements during sentence reading in adults (Plummer, Perea & Rayner, 2014).

Why should document count matter? One possibility is that document count is associated with semantic diversity, broadly defined as the extent to which different contexts are similar in meaning. Consistent with this idea, semantic diversity is closely associated with lexical decision performance in adults: words experienced in contexts that are semantically diverse are processed faster than words that are experienced in less semantically diverse contexts (Hoffman & Woollams, 2015; Jones, Johns & Recchia, 2012). At the same time, people are slower to make judgements about words high in semantic diversity in tasks that tap meaning. For example, Hoffman and Woollams reported slower response times in a word association task for words higher in semantic diversity, despite the same words being processed more quickly in lexical decision.

The effect of semantic diversity on word learning has also been examined. Jones et al. (2012) used an artificial learning paradigm with adults and found that increasing the number of exposures to a word influenced learning only when the repetitions were accompanied by a modulation in semantic context: merely repeating the episode without varying its semantic content did not influence learning. Similarly, Johns, Dye and Jones (2015) exposed adults to novel words in semantically diverse contexts or more uniform contexts. Greater diversity during learning was associated with faster recognition, as indexed by performance in a lexical decision task. Johns et al. also found

that the meanings of items trained in redundant contexts were better discriminated than those experienced in more diverse contexts. These results fit comfortably with the finding that words high in semantic diversity enjoy a processing advantage in lexical decision but are slowed in tasks that require semantic access (Hoffman & Woollams, 2015).

In contrast to these findings, Bolger, Balass, Landen and Perfetti (2008) found that contextual variation led to better learning of meaning. Adults read sentences containing rare, unfamiliar words, either in a novel sentence each time, or in the same sentence repeated (therefore a manipulation of contextual diversity rather than semantic diversity). At post-test, words experienced in multiple contexts showed an advantage in meaning generation and sentence completion tasks, consistent with their meaning having been better abstracted than words experienced in non-varying contexts, but contrasting with the findings discussed above. Why might this be? First, it could be due to differences in the kind of semantic knowledge needed to complete the post-tests. While Bolger et al. used a definition and a sentence completion task to measure semantic learning, Johns et al. (2015) used a semantic similarity judgement task in which participants rated how similar each new word was to existing words. Plausibly, these tasks tap word knowledge in very different ways, leading to the different pattern of results across the two experiments.

There are other important differences between the two experiments too. In Johns et al., participants read authentic discourse contexts containing a pseudoword that replaced a real word (e.g., *covella* for *constellation*) and across the course of the experiment they read 10 novel words. In Bolger et al., participants read rare words in single sentences, with 72 rare words being encountered over the course of the

experiment. In addition while learning was incidental in the Johns et al. study, participants were explicitly instructed to try to learn word meanings in Bolger et al. These methodological differences make it difficult to draw clear conclusions. Moreover, both experiments investigated word learning in undergraduate students, not children. One study has explored the question of contextual variation in children's word learning – but from spoken language experience, via storybook reading (Horst, Parsons, & Bryan, 2011). While word-object referent mapping was better following repeated than varying contexts, referent mapping does not capture the partial or incremental nature of word learning that characterises the gradual accumulation of word knowledge via independent reading episodes; and, like Bolger et al., contextual variation was captured by manipulating document count, rather than semantic diversity.

In summary, it is clear that contextual diversity affects word learning but the exact nature of the effect remains unclear and differs across tasks. Previous studies have used single word tasks to index word knowledge, and while informative, these tasks do not speak to how that knowledge is acquired, and what readers do with their emerging word knowledge when they encounter the partially-learned words in text. Furthermore, while stability and context-independence are thought to be hallmarks of high lexical quality (Perfetti, 2007), this cannot be the case for words with high semantic diversity: they are necessarily context-variable and a good understanding of the meaning of a high diversity word demands tolerance of context variability, perhaps explaining the longer time required for making semantic judgements. It is therefore important to investigate further both the process of acquiring new words in high and low diversity contexts, and the impact these two exposure types has on reading the encountered words in text.

Using eye movements to measure incidental word learning via reading

Joseph, Wonnacott, Forbes and Nation (2014) developed an incidental learning paradigm that capitalised on the idea that as novel words become more familiar, they should be fixated for a shorter time. Adults read a series of sentences containing novel words and their eye movements were monitored over repeated exposures. Reductions in reading times were evident before participants had good explicit knowledge of the words; this demonstrates utility as a more implicit measure, sensitive to partial or fragile knowledge (see also Elgort, Brysbaert, Stevens & Van Assche, in press). Alongside fixation duration on the novel words, the eye movement record provides more detailed information that informs our understanding of how people process novel words, as they encounter them in text. Skilled readers not only show longer first fixations on novel words (Chaffin, Morris & Seely, 2001; Lowell & Morris, 2014), they also make more regressions back to novel words than familiar words (Chaffin et al., 2001), suggesting that they use later contextual information to help work out a possible meaning for a newly encountered word.

Using eye movement measures within a learning paradigm offers a powerful means to tap the gradual accumulation of word knowledge, as readers experience novel words in text. By manipulating the nature of the repeated exposures, we can examine the conditions that best support learning, as indexed by subtle changes in eye movement behaviour. Furthermore, it enables us to measure reading in a natural and ecologically valid way, as children read connected text silently. Rather than bisecting behaviour using tasks that tap learning of form (e.g., lexical decision) or meaning (e.g. semantic similarity ratings), we can measure how emerging word knowledge influences reading behaviour, as children encounter novel words in text. This provides a measure of processing which is direct and implicit, rather than inferred from a secondary task

such as reading aloud, lexical decision or semantic classification, tasks that have their own metacognitive load and performance limitations.

The current experiment

Our aim is to test a key feature of the lexical legacy hypothesis (Nation, 2017), the idea that the linguistic environment a word appears in during its lexical history within an individual's experience influences its subsequent lexical quality. To test whether semantic diversity affects how well a word is learned, we exposed children to novel words in one of two conditions. In the non-diverse condition, each novel target word was presented in ten different sentences which shared a common semantic context. In the diverse condition, the same novel word appeared in semantically diverse contexts (see Table 2 for example sentences). The two conditions were identical in terms of frequency of exposure to the novel words, and the number of unique documents each was seen in. Critically however, they differed in semantic diversity. This allowed us to move beyond document count as a proxy for diversity and instead examine the effect of variations in semantic context on the accumulation of word knowledge from experience.

It is helpful to provide a brief overview of our design, before describing how learning was measured and what our key predictions were. Our to-be-learned target words were low frequency verbs, chosen to be unfamiliar to children. Verbs allowed us to move beyond the word-object referent mapping tasks that characterise word learning studies in infancy to consider words that have more complex and nuanced meanings. Using words rather than nonwords also added to the validity of our experiment, and allowed us to use authentic contexts. The experiment comprised three phases and took place over two sessions on consecutive days (see Figure 1). Pre-exposure, children

encountered the novel words in sentences that were either informative or neutral regarding the novel word's meaning (see Table 3). In the exposure phase, children read sentences each containing a novel word embedded in either diverse or non-diverse semantic contexts. Finally, in the post-exposure phase, they once again read novel words in informative and neutral sentences. Throughout all three phases, eye movements were monitored. Our aim was to track reading behaviour incrementally as word knowledge accumulated across the experiment and for the test phase to be indistinguishable from the exposure phase from the child's perspective. The children then completed three offline post-tests: spelling, to measure how well they had learned the orthographic form; and two tasks tapping the acquisition of meaning.

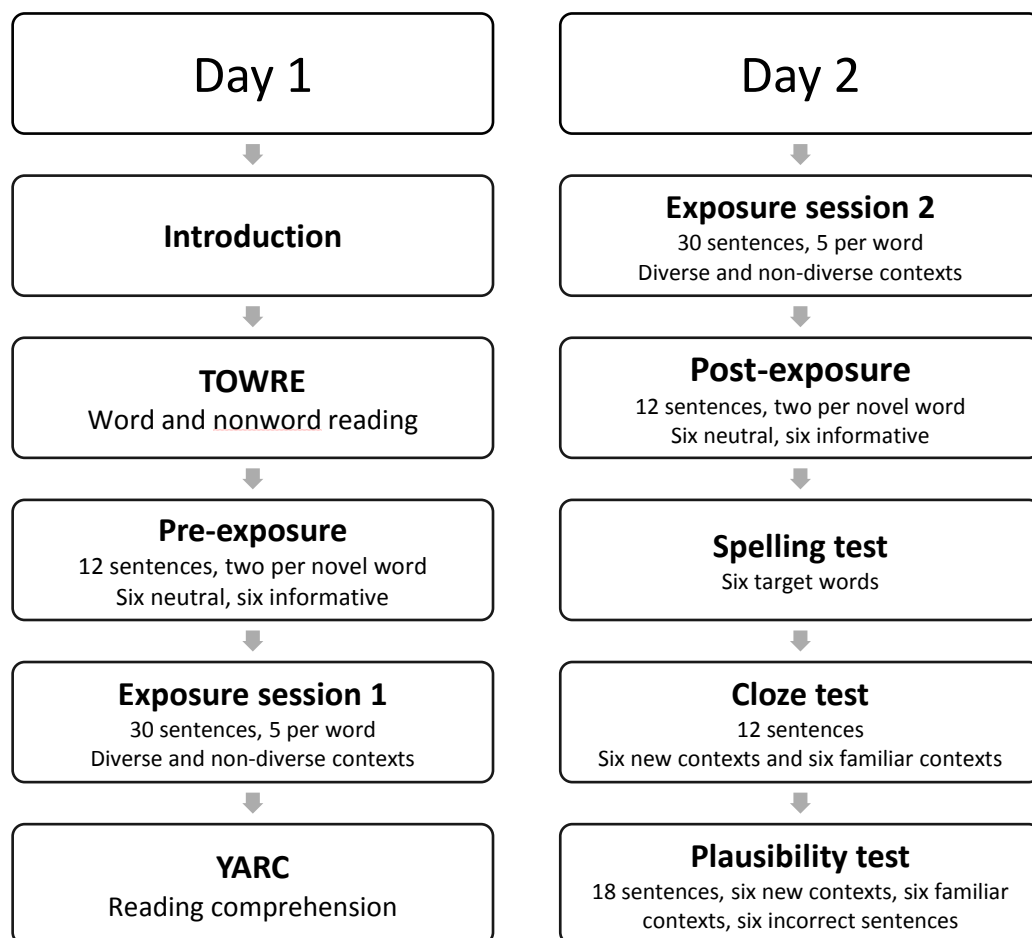


Figure 1. Outline of testing procedure

The acquisition of word knowledge was assessed in two ways: via performance on the three post-tests and via changes in reading behaviour as measured by eye movements. The post-tests tapped learning of form and learning of meaning, allowing us to compare our findings with those from studies of adults, discussed earlier (Bolger et al., 2008; Johns et al., 2015). We predicted better learning of form following diverse exposure, given the association between semantic diversity and lexical decision in skilled readers (Hoffman & Woollams, 2015; Jones et al., 2012). Thus, words experienced in diverse contexts should be more accurately spelled post-exposure. The effect of diversity on the two semantic tasks is harder to predict from the extant literature. Bolger et al. (2008) observed better learning of meaning for those words experienced in more diverse contexts. In other experiments, greater diversity is associated with poorer performance on semantic measures (Hoffman & Woollams, 2015; Johns et al., 2015). As our two semantic tasks require relatively deep semantic knowledge, we might expect results more akin to those of Bolger et al.

It is harder to make clear predictions about the eye movement data for two reasons. First, there is no clear distinction between indices of meaning vs. form when reading meaningful text. Second, despite the clear utility of the method, there has been no previous investigation of the conditions that influence incidental word learning from reading experience using eye movements. Most generally, we predicted that reading times would reduce with increasing exposure and familiarity with the novel words during the exposure phase. If the diversity of the context in which the novel words are encountered is implicated in incidental word learning, then reading times should not only be predicted by quantity of exposure, but also by our diversity manipulation. We thus expected shorter reading times on target words seen in diverse than non-diverse

contexts after exposure. The comparison between processing in the neutral and informative sentences read before the exposure phase vs. after the exposure phase provided an opportunity to assess how semantic diversity influences learning. Our predictions here centred on the idea that the product of increased familiarity with a new word is greater context-independence. Therefore, we would expect to see a reduction in reliance on context with increasing exposures, indexed by longer reading times on novel words in neutral than informative sentences before exposure (because children spend longer reading the contexts rather than the novel words in informative contexts) but little difference between informative and neutral (now redundant) contexts after exposure. Thus we predicted reading behaviour (reading times and regressions) in the informative and neutral contexts to be more similar after exposure than before exposure. In addition, as word knowledge for the words encountered in the semantically diverse contexts would be expected to be less context-specific, we might expect a smaller difference in reading times between informative and neutral contexts after as compared to before exposure specifically for those words encountered in diverse contexts during exposure.

Finally, we took the opportunity to explore individual differences in word learning. Children with poor reading comprehension are less able to use discourse context to infer the meaning of novel words (Cain, Oakhill & Elbro, 2003; Cain, Oakhill & Lemmon, 2004); they are also poor at learning the semantic attributes of novel objects (Nation, Snowling & Clarke, 2007). Similarly, in Bolger et al.'s (2008) experiment with adults, reading comprehension predicted how well the meanings of new words were learned. We therefore predicted that reading comprehension skill, as measured by a standardised test, would be associated with how well children were able to learn the

words in our experiment. Individual differences in reading comprehension have not been assessed with respect to eye movement behaviour in a word learning task, though poor comprehenders are less sensitive to cues from sentential context when reading words aloud (Nation & Snowling, 1998). Potentially therefore, differences in comprehension skill might be associated with differences in how sentences are read during the exposure phase.

Method

Participants

Forty-seven children in Years 5 and 6 were recruited from three local state primary schools. Informed consent was obtained from parents of all children who took part. Of the 47 recruited, seven were excluded: three due to tracker loss, one who had taken part in a pilot study, one whose word reading skill was below the normal range (see below for details), one who did not read any of the sentences in full, and one whose data were not recorded due to experimenter error. This left 40 children (27 female; mean age = 10.7 years, SD = 0.6) who met inclusion criteria and had full and usable data sets. Six children were bilingual, but all were fully fluent in English, having been educated in English and they had no difficulty completing the tasks. To establish that all children had sufficient word reading skills to complete the experiment, we screened our sample using the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999). This requires children to read aloud as many words and non-words as possible from a list in 45 seconds. As mentioned, only one child scored more than 1SD below the mean and so was excluded from further data collection. We used the York Assessment of Reading for Comprehension (YARC; Snowling et al., 2009) to measure

reading comprehension, and to provide an additional measure of reading accuracy and fluency. In this test, children are asked to read aloud two passages and then answer eight comprehension questions about each one. As can be seen from Table 1, our sample showed average to high-average reading skills. Children were randomly allocated to the diverse or the nondiverse group meaning that each child read all words in either the diverse or nondiverse contexts.

Table 1. Mean (SD) age and performance of the children on standardized tests of reading

	Year 5¹ (n=27)	Year 6¹ (n=13)
Age (years)	10.3 (0.3)	11.2 (0.5)
TOWRE ² words	110 (17)	115 (17)
TOWRE nonwords	113 (11)	116 (19)
YARC ³ accuracy	110 (9)	107 (11)
YARC reading rate	116 (12)	117 (14)
YARC comprehension	113 (13)	116 (11)

Notes. ¹Standard score M = 100, SD = 15. ²TOWRE (Test of Word Reading Efficiency; Torgeson et al., 1999). ³YARC (York Assessment of Reading for Comprehension; Snowling et al., 2009).

As our diversity manipulation was between participants, it was important to establish that the two groups did not differ on any variable that would be likely to affect their performance on our tasks. A series of t-tests (Table 2) showed that the two groups did not differ from one another on any measure of reading skill, nor in age, proportion of girls, proportion of bilingual speakers or mean first pass or total reading times.

Table 2. Mean age, reading scores, number of girls, number of bilingual speakers and mean first pass and total reading times for children in the diverse and nondiverse groups. Standard deviation are in parentheses.

	Diverse	Nondiverse	<i>t</i>	<i>p</i>
Age (years)	10.59 (0.66)	10.71 (0.54)	.62	<i>p</i> > .5
TOWRE Words (raw score)	81.35 (8.98)	78.00 (11.53)	1.0	<i>p</i> > .3
TOWRE Nonwords (raw score)	48.60 (10.56)	47.05 (8.09)	.51	<i>p</i> > .6
YARC Accuracy (ability score)	65.85 (7.42)	64.55 (5.80)	.60	<i>p</i> > .5
YARC Rate (ability score)	86.99 (8.09)	85.05 (9.57)	.68	<i>p</i> > .5
YARC Comprehension (ability score)	70.73 (7.53)	71.45 (9.96)	.27	<i>p</i> > .7
Number of girls (/20)	12	15		<i>p</i> > .5
Number of bilinguals (/20)	3	3		<i>p</i> = 1
Mean gaze duration (ms)	419 (236)	426 (248)	.13	<i>p</i> > .8
Mean total reading time (ms)	631 (379)	597 (382)	.52	<i>p</i> > .6

Materials.

(i) *To-be-learned target verbs.* We started with 14 low frequency past tense verbs. We chose the past tense as it is difficult to embed infinitive or present tense forms into multiple sentence frames. In addition, the -ed ending provided an additional clue that the target word was a verb. We then administered an online questionnaire to 29 children aged 8-12 years who did not take part in the main experiment. We used a relatively wide age range for the pre-screen in order to reveal when our target words became familiar. We asked the children to decide which of four categories best described their knowledge of each target word: (1) I've never seen it before; (2) It looks familiar but I don't know what it means; (3) I have an idea of its meaning; and (4) I definitely know its meaning. If they had chosen category (3) or (4), children were asked to type in a definition of the word. Scores were tallied and from this, we eliminated words with a mean score of 2.5 or above (3 words: *deteriorated*, *degenerated*, *disclosed*).

We also eliminated one word with a rare initial trigram (*thwarted*) and three more due to orthographic overlap with other candidate target words (*interceded*, *integrated*, *divulged*); one word was excluded as it was shorter than others (*impeded*). This left us with six target words (*accumulated*, *amalgamated*, *confabulated*, *exacerbated*, *intervened*, *languished*) which were not well known to the children and broadly comparable in word length (mean number of characters = 10.83, SD = 0.75), familiarity to children aged 8-12 (M familiarity score = 1.79, SD = .30), number of syllables (M = 4.17, SD = 1.33), number of phonemes (M = 10.50, SD = 2.07), number of morphemes (M = 3.33, SD = 0.82), and bigram frequency (M = 2472, SD = 615). These data were extracted from the English Lexicon Project database (Balota *et al.*, 2007).

(ii) *Sentence contexts for the exposure phase.* Each to-be-learned word was embedded in two sets (diverse and non-diverse) of 10 sentence frames which provided some information about target word meaning. Table 2 shows an example. Sentences in the two diversity conditions did not differ in length, and target words were never the first or the last word in a sentence. The sentences were created with the intention that the full meaning of the word was unlikely to be gained through a single exposure but that 10 exposures would be sufficient for children to start to build a rudimentary representation of its meaning. To provide a validity check on our diversity manipulation, we asked 32 adults to complete a short online questionnaire in which they read all ten sentences for each target word (with the target word itself removed) in either the diverse or nondiverse condition and rated how similar in topic the sentences were to one another on a scale of 1 (not at all similar) to 7 (extremely similar). Diverse sentences ($M = 4.13$, $SD = 1.65$) and nondiverse sentences ($M = 5.36$, $SD = 0.87$) were rated as significantly different from one another in similarity, $t(31) = 3.73$, $p < .001$.

The number of exposures was informed by a pilot study in which children received only six exposures and subsequently showed poor knowledge of word meanings at post-test. Semantic diversity was manipulated between participants such that each child read all six novel words either in diverse or in non-diverse contexts.

Table 3. Example stimuli. The to-be-learned target in this example is accumulated (not shown in bold in the experiment). A full list of experimental sentences can be found in the Appendix.

	Non-diverse context (law/evidence)	Diverse context
Experimental sentences	<p>Enough proof had accumulated so that the jury could make a fair judgement on the case.</p> <p>The police accumulated a lot of strong evidence which meant they could arrest the thief.</p> <p>Members of MI5 accumulated all the incoming data and saved it onto a computer file.</p> <p>After the news report went out, the police accumulated more than 25 witnesses.</p> <p>The lawyer accumulated witness statements to get support for the case.</p> <p>The burglar accumulated information about the neighbourhood before committing the crime.</p> <p>The evidence accumulated until there was no question that he was guilty.</p> <p>The proof that she had stolen the money accumulated over time and eventually she lost her job.</p> <p>The witness statements accumulated and in the end he decided to plead guilty.</p> <p>The solicitor accumulated the documents for the case and took them to court.</p>	<p>Enough proof had accumulated so that the jury could make a fair judgement on the case.</p> <p>The woman forgot to clean under the bed, so dust had accumulated on the floorboards.</p> <p>The girl loved collecting rubbers and accumulated more each week using her pocket money.</p> <p>After just one week at his new school, the boy had already accumulated several new friends.</p> <p>The doctors accumulated enough test results to diagnose and treat the patient.</p> <p>Lava had accumulated beneath the surface which caused a spectacular eruption from the volcano.</p> <p>His debts accumulated until he had to sell his house to pay off the loan.</p> <p>Although she had accumulated a lot of wealth, this meant she also had to pay a lot of tax.</p> <p>She was shocked to discover how many emails had accumulated while she was away.</p> <p>The fluid had accumulated in his lungs and he found it very hard to breathe.</p>
Pre and	<p>Neutral</p> <p>The fisherman accumulated many stories.</p> <p>The children accumulated five apples.</p>	

post-test
sentences

Informative

The detective had accumulated enough evidence to catch the criminal.

The burglar had accumulated many stolen items over the years.

(iii) *Sentences for the pre- and post-exposure phases.* Four sentence frames were created for each target word: two neutral and two informative (see Table 3). The neutral sentence frame offered relatively little information about the meaning of the word, while the informative sentence frame provided a similar amount of information about the word's meaning as the sentences constructed for the exposure phase. The informative sentence frame always utilised the same semantic context as the exposure sentences in the non-diverse condition (e.g. a legal context for *accumulated*; see Table 3). Children read one neutral and one informative sentence for each target word at pre-exposure and the other neutral and informative sentences at post-exposure (counterbalanced across participants). Neutral sentences were shorter than informative sentences.

Apparatus

Throughout pre-exposure, exposure, and post-exposure, children's eye movements were recorded using an Eyelink 1000 eye tracker (SR Research; Mississauga, Canada) as they read sentences from a 14" computer monitor at a viewing distance of 62 cm. Sentences were presented in a white, monospaced font (Consolas), size 14, on a black background. Eye movements were monitored at a rate of 1000 Hz. Although the children read binocularly, only the movements of the right eye were monitored.

Procedure

The testing schedule is summarised in Figure 1. Testing took place in a quiet area close to the child's classroom. On Day 1, children completed the TOWRE, the pre-

exposure, the first exposure session and the YARC. On Day 2, children completed the second exposure session, the post- exposure, and three offline post-tests.

For the eye tracking components, children sat in a customised chair in front of a computer monitor, supported by a chin rest and a forehead rest to ensure comfort and to minimise head movements. They first undertook a calibration procedure during which they looked at each of three fixation points extending horizontally across the centre of the computer monitor. Following calibration, two practice trials were presented immediately followed by the experimental sentences. Each sentence was preceded by the appearance of a small fixation square. Children were instructed to fixate the square in order to trigger the appearance of the sentence, thus ensuring accuracy for each trial. After reading, children pressed a button on a handheld gamepad controller to terminate the trial. If the child did not press the button within 30 seconds of the text appearing, the display was automatically terminated. Calibration accuracy was checked following each sentence and the tracker was recalibrated if necessary (maximum calibration error was set at 0.5°). There was a brief break halfway through each exposure session but not during the pre/post-exposure sessions as these were short.

(i) Sentence reading during pre-exposure, exposure and post-exposure. To ensure that the neutral sentence frame was always read before the informative sentence frame for each target word (so that clues about word meaning inferred from the informative context would not influence reading times on the target words in the neutral context), the order of presentation was the same for all participants, with the neutral sentence frame always immediately preceding the informative sentence frame for each word. Two practice trials always preceded the experimental trials.

During each exposure session, children read 30 sentences per day: 5 for each target word. This meant that over the two sessions, each of the six target words appeared in 10 different sentences in total. Each sentence was read once and sentences were presented in a random order for each participant. Two practice sentences and six filler sentences were also presented in each session, with fillers interleaved with the experimental sentences. Fillers contained a target word which was relatively low in frequency but which children were likely to know well (e.g. *fascinating*, *experimenter*). To encourage the children to read carefully, they responded to yes/no comprehension questions on a gamepad following one of the practice trials and all six filler sentences. There were no questions relating to the experimental trials in order to prevent an additional learning opportunity for some of the target words. Mean accuracy for comprehension questions was 86%.

(ii) Offline post-tests. On Day 2, immediately after reading the post-exposure sentences, children completed three pen-and-paper post-tests to assess their learning of the target words. The first was a spelling test to measure learning of orthographic form. Children were simply asked to write all six target words. The order in which they were asked to write them was counterbalanced across participants. Each word was awarded a total of two marks if there were no errors, one mark if there were one or two letters that were incorrect (including omissions, transpositions and additions), and no marks if more than two letters were incorrect.

Next the children completed a written cloze task. They were given the six target words and were asked to use these to complete 12 sentences. For each target word, one sentence had the same semantic context as the non-diverse condition (e.g. legal theme for *accumulated*) and one sentence had a new semantic context that the children had

not encountered during exposure (e.g. *'The garden pond hadn't been cleaned for months and pondweed had _____' for accumulated*). This was to examine whether children could generalise what they had learned during exposure to a new context.

Finally children completed a plausibility task in which they read 18 sentences all containing a target word and had to indicate with a tick or a cross whether or not the sentence made sense. There were three types of sentence: (i) correct sentences in the same semantic context as the non-diverse condition (e.g. *'The judge had accumulated all the evidence so he could make a decision'*); (ii) correct sentences in a new semantic context that had not been encountered during exposure (e.g. *'People from all over the city had accumulated in the central plaza'*); and (iii) sentences which did not make sense (e.g. *'Sarah had accumulated her left knee quite badly when she fell off her bike.'*). Each correct answer was awarded one mark.

Following the three post-tests, children were thanked for their time and given a certificate and sticker.

Results

Data were analysed in the R computing environment (R Development Core Team, 2012) using Linear Mixed Effects models (Baayen, 2008; Jaeger, 2008). All models included random intercepts for participants and items and random by-participant and by-item slopes for all fixed effects (i.e. a full random slopes structure, see Barr, Levy, Scheepers, & Tilly, 2013). When a model did not converge (mostly models including comprehension skill as a fixed effect), we first took out interactions between random slopes and then removed random slopes one by one (removing those

that accounted for the least variance) until the model converged. Note that this was rarely necessary for reading time data; however, it was usually necessary to remove random slopes in order to achieve model convergence for binary outcome variables including regression probabilities and the two semantic post-tests). We report only the converged models.

We centred all fixed effects using contrast coding to reduce the effects of collinearity between the main effects and interactions and in order that main effects were evaluated as the average effects over levels of the other predictors. Regression coefficients, standard errors (SE) and t (for reading time measures) or z (for regression probabilities) values are reported. Following Vorstius et al. (2013), we used the two-tailed criterion (t or $z \geq 1.96$ SE), corresponding to a 5% error criterion for significance for all tests, but with an adjustment for multiple comparisons. von der Malsburg and Angele (2017) argued that although eye movement researchers tend not to make adjustments for multiple comparisons, doing so reduces false positives and does not result in a reduction of statistical power to an unacceptable level. Therefore we made appropriate adjustments (dividing the α threshold by the number of eye movement measures or post-tests while accounting for the average correlation between measures or scores) and treated as significant only those effects that reached this threshold. Note that although unconventional within the field and conservative in terms of the likelihood of committing a Type I error, we agree with Simmons, Nelson and Simonsohn (2011) that false positives are potentially more harmful to scientific discourse than false negatives (Type II errors).

For all eye movement data, fixations shorter than 80ms and longer than 1200ms were excluded and trials which showed blinks or tracker loss on the target word were

also deleted. As is common in eye movement studies reading time data were not normally distributed and so were log transformed. We also deleted outliers (more than 2.5SD from the mean) for all reading time measures. This resulted in the removal of 5.8% of the eye movement data across both exposure and pre- and post-exposure sessions.

1. How well did the children learn the target words?

We begin by describing the results from the three offline post-tests. Our main questions were whether children learned the target words, whether semantic diversity during exposure affected how well children performed in the three tests, and whether they could generalise the word meaning to a new context in the two semantic tests.

Descriptive data are shown in Figure 2.

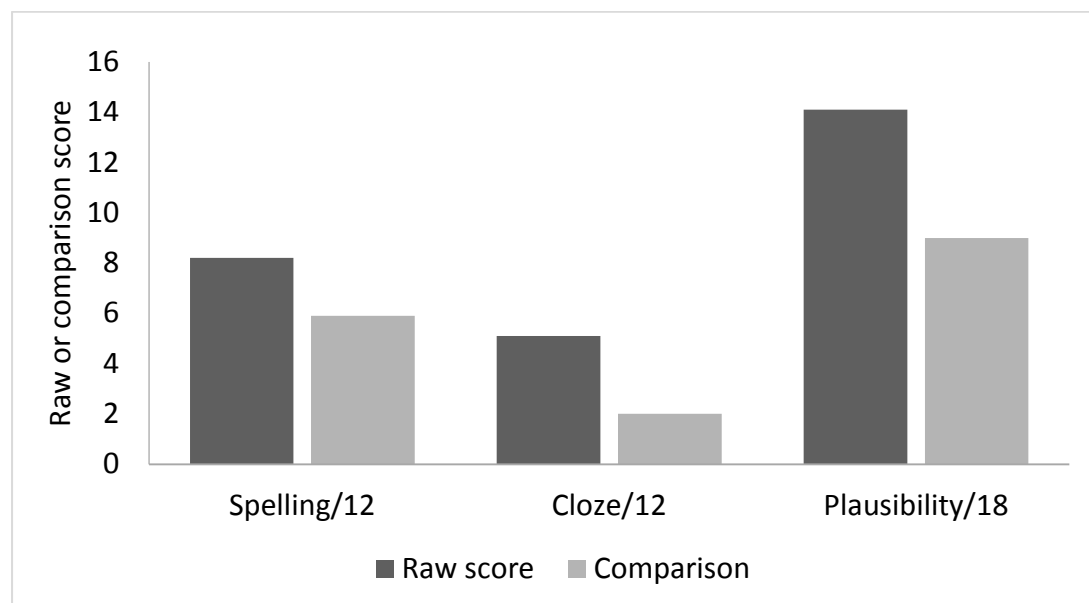


Figure 2: Obtained and comparison scores on the spelling, cloze and plausibility post-tests. Obtained scores are mean raw scores obtained by participants in the experiment. Comparison scores are scores that would be obtained by chance for the cloze and

plausibility tests. For the spelling test, comparison scores are those obtained by 38 children who did not take part in the main experiment.

Table 4 shows the proportion of each response type (0, 1 or 2) in the spelling task for words experienced in diverse and nondiverse contexts. To assess learning of orthographic form, we first compared spelling accuracy at post-test relative to baseline data, provided by the 38 children who did not participate in the main experiment and thus had not been exposed to the target verbs. Spelling was more accurate post-exposure than at baseline ($M = 8.2$ vs. 5.9), indicating that the new forms had been learned, $t(76) = 5.06$, $p < .001$. We ran a model with contextual diversity and comprehension skill as fixed effects and participant and word as random effects to examine the effect of semantic diversity in the spelling post-test (see Table 5). The only reliable effect was that of comprehension skill with better comprehenders spelling more words correctly. There was no effect of diversity, indicating equivalent learning in both conditions (Diverse: $M = 8.40$, $SD = 2.04$; Non-diverse: $M = 7.90$, $SD = 1.90$).

Table 4: Proportion of response type (0, 1 or 2 points) in the spelling post-test for words encountered in diverse and non-diverse conditions. Standard deviations in parentheses.

	Score	Diverse	Nondiverse
Spelling test	0	0.13 (0.13)	0.15 (0.16)
	1	0.35 (0.24)	0.38 (0.21)
	2	0.53 (0.27)	0.48 (0.22)

Table 5: Results of models examining the effect of contextual diversity, context and comprehension skill on offline test performance

	Spelling	Cloze	Plausibility
Diversity (diverse vs. non-diverse)	$b = .08, SE = .11, t = 0.73$	$b = .20, SE = .40, z = 0.50$	$b = .25, SE = .29, z = 0.86$
Comprehension (YARC score)	$b = .01, SE = .01, t = 2.46^*$	$b = .10, SE = .02, z = 4.16^*$	$b = .07, SE = .02, z = 4.12^*$
Diversity * Comprehension	$b = .02, SE = .01, t = 1.57$	$b = .03, SE = .05, z = 0.53$	$b = .001, SE = .03, z = 0.20$
Context (Old vs. New)		$b = .34, SE = .22, z = 1.53$	$b = .13, SE = .26, z = 0.50$
Diversity * Context		$b = .59, SE = .45, z = 1.33$	$b = .99, SE = .51, z = 1.94$
Context * Comprehension		$b = .02, SE = .03, z = 0.68$	$b = .02, SE = .03, z = 0.75$
Context * Comprehension * Diversity		$b = .05, SE = .06, z = 0.91$	$b = .04, SE = .06, z = 0.70$

*two-tailed significance criterion (t or $z \geq 2.24$), corresponding to a 5% error, adjusted for multiple comparisons.

Turning to the two measures of semantic learning, cloze and plausibility, both contained two context types. One was the same as the non-diverse context during exposure. Note that this was seen in both exposure conditions, but nine times more for words encountered in the non-diverse condition than the diverse condition. The other context was new and had not been encountered in either exposure condition. For these two tests, we did not obtain baseline data from children who did not take part in the main experiment as this would have been extremely demotivating as children were very likely to perform at floor. Instead we compared performance of children in our experiment to chance performance. Data for the two context types are shown in Figure 2 and the results of the analyses are summarised in Table 5. There is evidence that semantic learning happened. In the cloze task, if children had randomly picked each word twice, chance performance would be 2/12; the mean score was 5.1. In the plausibility test, performance was significantly above chance, $t(39) = 9.91, p < .001$, showing that some semantic information had been learned. Given this evidence of learning, our next question was whether seeing a word in diverse contexts during exposure better equipped a child to understand it in a novel context (see Table 6). We ran two models to test this, one for each semantic post-test in which we entered semantic diversity, novelty of context, and comprehension skill as fixed effects, and participant and word as random effects. The only effects observed were of comprehension skill: children with stronger comprehension skills performed better on both measures (see Table 5).

Table 6: Proportion of correct responses in the cloze and plausibility post-tests for words encountered in old (previously encountered) and new (not encountered) semantic contexts in diverse and non-diverse conditions. Standard deviations in parentheses.

	Old context		New context	
	Diverse	Non-diverse	Diverse	Non-diverse
Cloze test	0.41 (0.49)	0.50 (0.50)	0.40 (0.49)	0.39 (0.49)
Plausibility test	0.82 (0.381)	0.77 (0.42)	0.73 (0.44)	0.84 (0.37)

2. How did the children read the target words? Comparing pre-and post-exposure reading times and regressions

To examine the effect of exposure, semantic diversity, contextual informativeness and comprehension skill on incidental word learning, we ran one model for each eye movement measure of interest. Based on previous studies (e.g., Chaffin et al., 2001), we selected five eye movement measures: *gaze durations* (the sum of all first pass fixations made on the target); *go past times* (the sum of all temporally contiguous fixations until the point of fixation progresses to the region to the right); *total reading time* (the sum of all fixations made on the target); *regressions out* (the probability of making a leftward eye movement out of a region before leaving that region to the right); and *regressions in* (the probability of making a leftward eye movement into a region having already left that region to the right). All models had four fixed effects: exposure (pre versus post), diversity (diverse versus non-diverse), context informativeness (neutral versus informative), and comprehension skill (continuous measure: ability score on the YARC) and random effects of participant and word. For the two regression measures there were problems with model convergence and so we ran these models without

comprehension skill as a fixed effect. Table 7 shows mean reading times and regression probabilities for target words at pre- and post-exposure in neutral and informative contexts and Table 8 shows the results of the models.

Table 7: Mean reading times and regression probabilities on the target word at pre- and post-test in neutral and informative contexts.

Reading time measures are in milliseconds. Standard deviations are in parentheses.

Diversity	Eye movement measure	Neutral context		Informative context	
		Pre-test	Post-test	Pre-test	Post-test
Diverse context	Gaze duration	466 (293)	399 (185)	438 (258)	378 (194)
	Go past time	732 (410)	486 (262)	644 (403)	478 (295)
	Regressions out	0.25 (0.43)	0.08 (0.28)	0.17 (0.38)	0.13 (0.34)
	Regressions in	0.39 (0.49)	0.31 (0.46)	0.18 (0.38)	0.16 (0.37)
	Total reading time	808 (405)	573 (324)	647 (389)	501 (326)
Non- diverse context	Gaze duration	478 (284)	378 (212)	461 (262)	390 (220)
	Go past time	726 (420)	524 (328)	585 (370)	486 (321)
	Regressions out	0.27 (0.45)	0.25 (0.44)	0.15 (0.36)	0.12 (0.32)
	Regressions in	0.33 (0.47)	0.32 (0.47)	0.19 (0.39)	0.17 (0.37)
	Total reading time	786 (458)	538 (329)	624 (369)	473 (309)

Table 8: Results of models examining the effect of contextual diversity, exposure, context informativeness and comprehension skill on all eye movement measures in the pre- and post-test analyses

		Comprehension (YARC score)	Exposure (pre- vs. post-test)	Context (neutral vs. informative)	Diversity (diverse vs. non-diverse)
Main effects	Gaze duration	$b = .004, SE = .005, t = 0.84$	$b = .13, SE = .04, t = 3.75^*$	$b = .01, SE = .07, t = 0.11$	$b = .05, SE = .08, t = 0.60$
	Go past time	$b = .01, SE = .01, t = 2.66^*$	$b = .31, SE = .03, t = 9.64^*$	$b = .12, SE = .10, t = 1.19$	$b = .01, SE = .12, t = 0.10$
	Regressions out		$b = .55, SE = .19, z = 2.82^*$	$b = .45, SE = .19, z = 2.29$	$b = .39, SE = .32, z = 1.22$
	Regressions in		$b = .17, SE = .16, z = 1.05$	$b = .92, SE = .16, z = 5.62^*$	$b = .07, SE = .24, z = 0.29$
	Total reading time	$b = .01, SE = .01, t = 1.97$	$b = .33, SE = .03, t = 10.58^*$	$b = .20, SE = .08, t = 2.64^*$	$b = .01, SE = .10, t = 0.08$
		Exposure * Comprehension	Diversity * Comprehension	Context * Comprehension	Exposure * Context
Interactions	Gaze duration	$b = .01, SE = .004, t = 2.86^{*a}$	$b = .003, SE = .01, t = 0.26$	$b = .004, SE = .004, t = 0.98$	$b = .01, SE = .07, t = 0.14$
	Go past time	$b = .01, SE = .004, t = 3.35^{*a}$	$b = .005, SE = .004, t = 1.06$	$b = .001, SE = .004, t = 0.31$	$b = .11, SE = .06, t = 1.76$
	Regressions out				$b = .45, SE = .39, z = 1.17$
	Regressions in				$b = .12, SE = .33, z = 0.37$
	Total reading time	$b = .01, SE = .004, t = 3.04^{*a}$	$b = .02, SE = .01, t = 1.43$	$b = .004, SE = .004, t = 0.93$	$b = .09, SE = .06, t = 1.40$
		Exposure * Diversity	Context * Diversity	Diversity * Exposure *Context	Diversity * Context * Comprehension
Interactions	Gaze duration	$b = .06, SE = .07, t = 0.90$	$b = .06, SE = .07, t = 0.80$	$b = .07, SE = .14, t = 0.47$	$b = .01, SE = .01, t = 1.48$
	Go past time	$b = .06, SE = .06, t = 0.86$	$b = .04, SE = .08, t = 0.52$	$b = .02, SE = .13, t = 0.17$	$b = .02, SE = .01, t = 1.94$
	Regressions out	$b = .60, SE = .39, z = 1.56$	$b = .92, SE = .39, z = 2.38^{*b}$	$b = 1.24, SE = .77, z = 1.61$	

	Regressions in	$b = .12, SE = .33, z = 0.38$	$b = .14, SE = .33, z = 0.44$	$b = .40, SE = .65, z = 0.62$	
	Total reading time	$b = .04, SE = .06, t = 0.59$	$b = .02, SE = .08, t = 0.21$	$b = .04, SE = .13, t = 0.30$	$b = .01, SE = .01, t = 0.90$
		Diversity * Exposure * Comprehension	Exposure * Context * Comprehension	Diversity * Exposure * Context * Comprehension	
Interactions	Gaze duration	$b = .004, SE = .008, t = 0.52$	$b < .002, SE = .008, t = 0.24$	$b = .01, SE = .02, t = 0.79$	
	Go past time	$b = .01, SE = .01, t = 1.89$	$b = .01, SE = .01, t = 0.98$	$b = .01, SE = .02, t = 0.84$	
	Regressions out				
	Regressions in				
	Total reading time	$b = .01, SE = .01, t = 1.52$	$b = .004, SE = .01, t = 0.52$	$b = .01, SE = .01, t = 0.52$	

**two-tailed significance criterion (t or $z \geq 2.38$), corresponding to a 5% error, with adjustment for multiple comparisons.*

^a In gaze durations, subset analyses showed no effect of comprehension skill at pre-test, $t < 1$, but a significant effect at post-test, $b = .01, SE = .005, t = 2.27$. In go past times, again there was no effect of comprehension skill at pre-test, $t < 1.8$, but a significant effect at post-test, $b = .02, SE = .005, t = 3.78$. We saw the same pattern in total reading times: no effect at pre-test, $t = 1.0$, but a significant effect at post-test, $b = .02, SE = .006, t = 3.38$.

^b Subset analyses showed no effect of context informativeness in diverse contexts, $b = .24, SE = .47, z = 0.52$, but a significant effect in nondiverse contexts, $b = 1.00, SE = .32, z = 3.17$, with more regressions out of the target word in the neutral than informative contexts.

There was an overall effect of comprehension skill in go past times: children with lower levels of comprehension skill showed longer reading times. For the comparison between pre-and post-exposure, there was a reliable effect in all reading time measures, with shorter reading times post exposure. There were also fewer regressions made out of the target word following exposure. The effect of contextual informativeness was evident in total reading times and regressions into the novel word, with longer reading times and more regressions in the neutral context, but this did not interact with exposure as predicted. There were no reliable effects of semantic diversity in any reading time measure.

Comprehension skill interacted with exposure phase in all reading time measures. Subset analyses, which examined the effect of comprehension skill at pre- and post-test separately, showed no effect of comprehension skill at pre-exposure but a reliable effect at post-exposure in all three measures, showing a greater reduction in reading times for children with better reading comprehension.

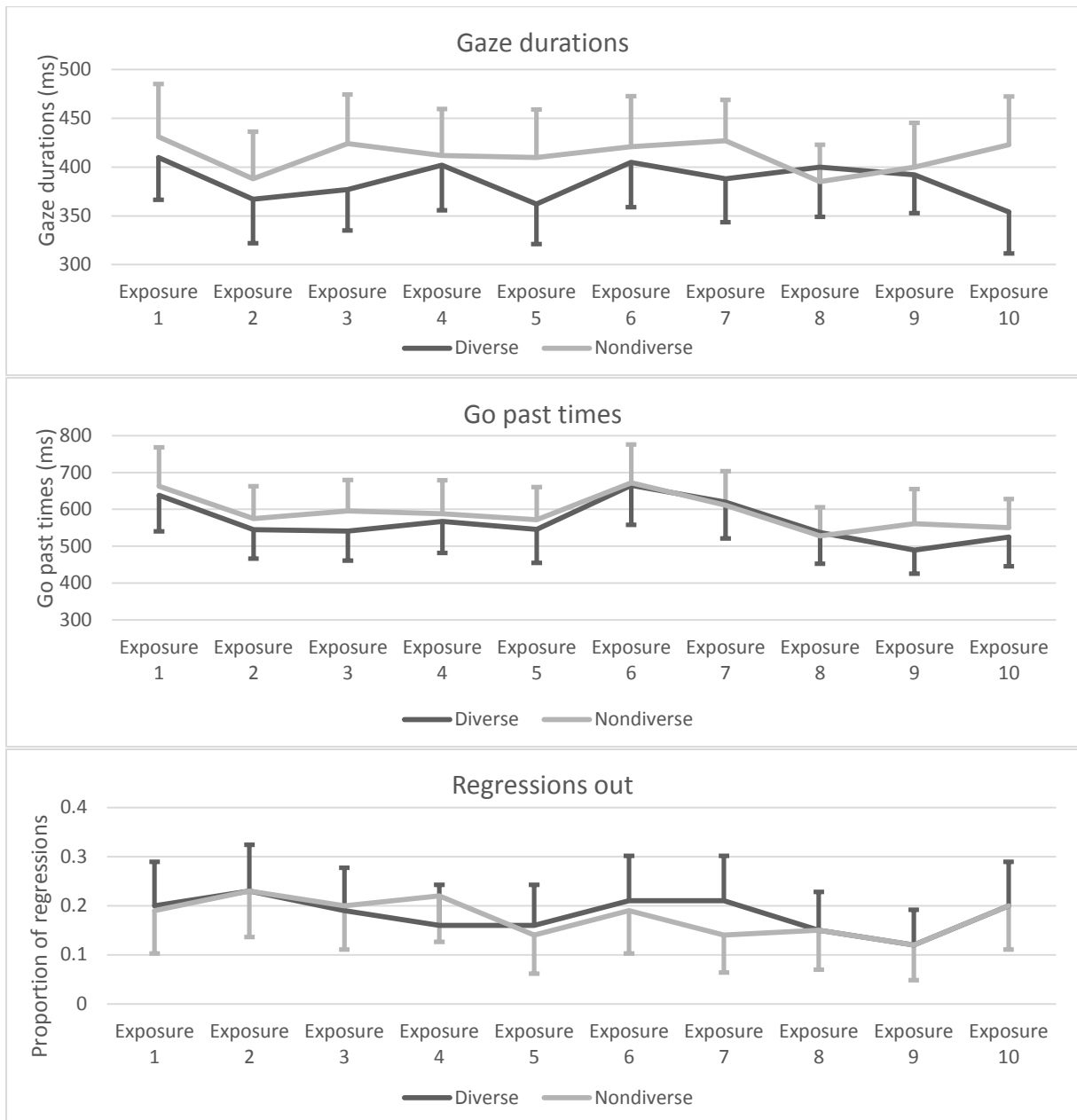
Finally, in the number of regressions made out of the target word, there was an interaction between diversity and context informativeness: children in the nondiverse group, but not the diverse group, made more regressions in the neutral than informative context.

Overall, we saw strong evidence that children learnt something about the novel words over the course of exposure with large decreases in reading times and regression probabilities in the post-exposure phase, particularly for children with better reading comprehension. There was also clear evidence that context informativeness plays an important role in children's reading behaviour with longer reading times and more regressions when the context is neutral. Overall, the nature of the exposure – whether words were seen in semantically diverse or non-diverse contexts – did not influence reading behaviour. There was however an interaction between diversity and context informativeness, indicating that children may have made more of an attempt to infer novel word meanings in neutral contexts (which are more challenging than in informative contexts) when they had experienced them in contexts that were semantically similar to one another rather than diverse.

3. How did the children read the target words during the exposure phase?

We now turn to reading times during the exposure phase to examine whether words encountered in diverse contexts were processed differently to those encountered in non-diverse contexts over time. These data are plotted in Figure 3. We included fixed effects of exposure (exposure sentence number, 1-10), diversity (diverse versus non-diverse) and comprehension skill (ability score on the YARC) and random effects of participant and word. Table 9 shows the results for these models. Note that for brevity,

we do not report all subset analyses for each individual exposure but as Figure 3 shows, in general there is a downwards trend over the course of exposure in all measures other than gaze duration (i.e. a reduction in reading times and regression probability).



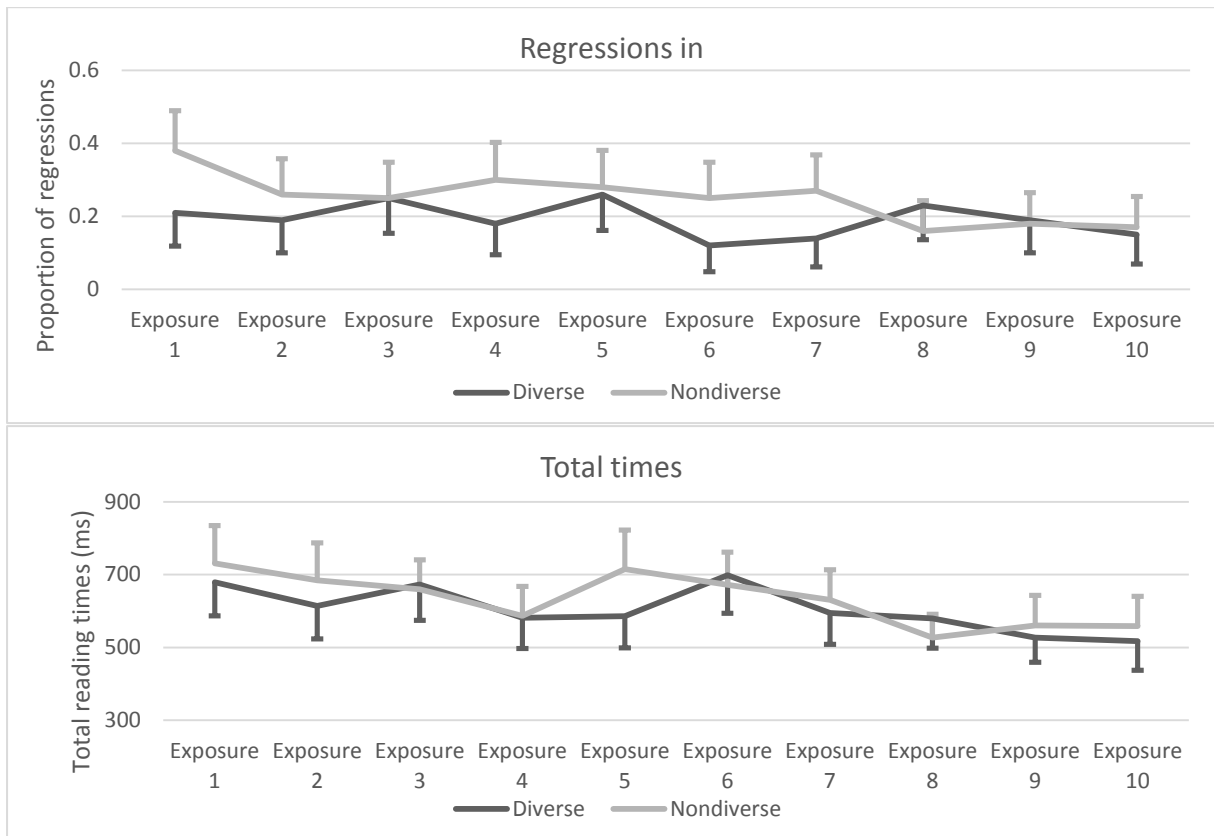


Figure 3. Reading times and regression probabilities on target words in diverse and non-diverse contexts across ten exposures.

Table 9: Results of models examining the effect of contextual diversity, exposure, and comprehension skill on all eye movement measures in exposure trials

		Comprehension (YARC score)	Exposure (1-10)	Diversity (diverse vs. non-diverse)		
Main effects	Gaze duration	$b = .01, SE = .004, t = 1.83$	$b = .002, SE = .003, t = 0.59$	$b = .06, SE = .06, t = 0.90$		
	Go past time	$b = .04, SE = .01, t = 5.27^*$	$b = .01, SE = .01, t = 2.37$	$b = .07, SE = .13, t = 0.49$		
	Regressions out	$b = .02, SE = .01, z = 1.58$	$b = .04, SE = .02, z = 1.92$	$b = .05, SE = .19, z = 0.25$		
	Regressions in	$b = .01, SE = .01, z = 0.48$	$b = .07, SE = .02, z = 4.01^*$	$b = .33, SE = .24, z = 1.41$		
	Total reading time	$b = .04, SE = .01, t = 4.69^*$	$b = .03, SE = .01, t = 4.96^*$	$b = .07, SE = .16, t = 0.47$		
		Exposure * Comprehension	Diversity * Comprehension	Exposure * Diversity	Exposure * Diversity* Comprehension	
Interactions	Gaze duration	$b = .001, SE < .001, t = 1.71$	$b < .001, SE = .01, t = 0.03$	$b = .01, SE = .01, t = 0.90$	$b < .001, SE < .001, t = 0.18$	
	Go past time	$b < .001, SE < .001, t = 1.05$	$b = .06, SE = .01, t = 4.17^{*a}$	$b < .001, SE = .01, t = 0.18$	$b < .002, SE < .001, t = 2.53^{*b}$	
	Regressions out	$b = .002, SE = .002, z = 0.76$	$b = .06, SE = .02, z = 3.00^{*a}$	$b = .004, SE = .04, z = 0.11$	$b = .01, SE = .004, z = 1.59$	

Regressions in	$b = .002, SE = .002, z = 1.12$	$b = .06, SE = .03, z = 2.25$	$b = .07, SE = .04, z = 1.89$	$b = .01, SE = .004, z = 0.25$
Total reading time	$b = .001, SE < .001, t = 2.09$	$b = .06, SE = .002, t = 4.02^{*a}$	$b = .01, SE = .01, t = 0.88$	$b = .001, SE = .004, t = 0.75$

**two-tailed significance criterion (t or $z \geq 2.38$), corresponding to a 5% error, with adjustment for multiple comparisons.*

^a In go past times, subset analyses showed no effect of comprehension skill in diverse contexts ($t = 1$) but an effect of comprehension skill in non-diverse contexts, $b = .02, SE = .009, t = 2.60$. In regressions, again subset analyses showed no effect of comprehension skill in the diverse condition ($z < 1$), but a significant effect in the nondiverse condition, $b = .04, SE = .009, z = 4.65$. Finally, in total reading times, subset analyses also showed no effect of comprehension skill in diverse contexts ($t = 1$) but an effect of comprehension skill in non-diverse contexts, $b = .02, SE = .01, t = 2.26$.

^b Subset analyses showed an interaction between exposure and comprehension skill in the diverse condition ($b = .001, SE = .001, t = 2.20$): while reading times decreased over the course of exposure for those with better comprehension skills, this was not the case for those with poorer comprehension skills. There was no such interaction in the nondiverse condition ($t < 1.2$).

Mirroring the findings discussed above, we saw an effect of comprehension skill in go past times and total times, with better comprehenders showing shorter reading times. There was an effect of exposure number in total reading times and regressions into the target word, both reducing with more exposures. There was no effect of diversity in any measure. We did however see an interaction between diversity and comprehension skill in go past times, regressions out of the target word, and total times: in all cases poorer comprehenders showed longer reading times or made more regressions in the non-diverse condition, but not in the diverse condition. Finally, there was a three way interaction between exposure, diversity and comprehension skill: in the diverse condition reading times decreased over exposure for those with good comprehension skills but not for those with poorer comprehension skills but this was not the case in the nondiverse condition (see Figure 4).

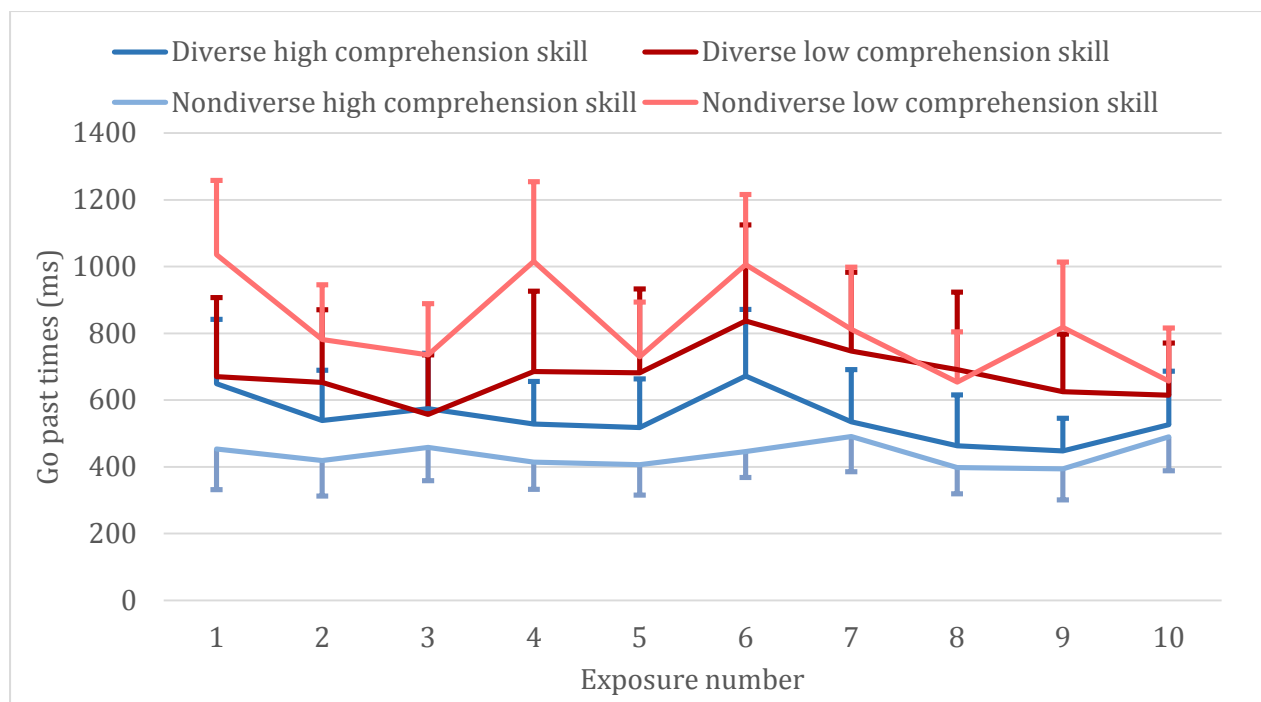


Figure 4. Mean go past times across ten exposures in diverse and nondiverse conditions for children 0.5SD s above and below the mean for reading comprehension skill. Error bars represent standard error.

Discussion

We used a novel approach to index incremental changes in moment-to-moment reading behaviour as children experienced novel words in ten different sentences. This allowed us to investigate whether contextual factors influence incidental word learning via reading. These online measures were complemented by more traditional pen-and-paper tasks that measured children's knowledge of the spelling and meaning of the new words. We also asked whether individual differences in reading comprehension skill were associated with word learning and the factors that influence word learning.

The first issue to address is whether our paradigm was sensitive to incidental learning. Put simply, how well did the children learn the words? Our data clearly show

that children learnt something about both the meaning of each new word, and its orthographic form. Across the ten exposures, children showed a reduction in total reading times on, and regressions into, new words and at post-exposure they showed a reduction in all reading time measures and in the number of regressions made out of the target words, compared to pre-exposure. Children were also better able to spell the novel words than children who had not participated in the exposure phase of the experiment. Thus, the children had formed a relatively well-specified and accurate orthographic representation of the target words. This is unsurprising given previous demonstrations of some orthographic learning following a single exposure to short and fairly concrete nouns (Nation et al., 2007; Share, 2004), but does show that longer verbs and more complex words (in terms of spelling-sound mappings) can be successfully encoded and reproduced after relatively few incidental encounters. The results also contribute to the now large literature showing the importance of print exposure for developing an orthographic lexicon (e.g. Mol & Bus, 2011).

Children were also able to infer word meanings from these few exposures with above chance performance in both semantic post-tests. To perform adequately in the two semantic post-tests children were required to abstract a core meaning across the different exposures. The superior performance in the forced choice plausibility task compared to the semi-productive cloze task suggests that although children's word knowledge was developing, it was not always sufficient to support using the new words productively. It is likely that more exposures would be needed for this to occur. We also made learning difficult as we did not test children until after all exposures. Previous research has shown that testing between learning trials has a positive effect on retention (e.g. Karpicke & Roediger, 2007); having children produce target words or

testing them after each exposure sessions might have enhanced learning and so increased performance in final post-tests. Nevertheless, children were good at recognising when sentences were implausible, showing good recognition and some understanding of the meaning of the new words.

Comparing reading times on the novel words in neutral vs. informative sentences in the post-exposure phase (relative to the pre-exposure phase) provided an opportunity to examine how context-independent the word had become. If via exposure children develop a well-specified and less context bound representation (Perfetti, 2007), we would expect less reliance on contextual information and hence little difference in reading times between words encountered in informative as compared to neutral sentences. Although we found longer reading times and more regressions in the neutral than informative context, we did not observe this predicted interaction, suggesting that children's lexical representations were not yet sufficiently well-established to be less context-bound. We would expect that with further exposures our predicted interaction would emerge, while acknowledging that effects of predictability will always play a role in lexical access during text reading.

It is clear that children learned something about the form and meaning of the novel words and that changes in online reading behaviour emerged as a function of exposure. It seems therefore that our paradigm offers a laboratory analogue of how children might construct word knowledge as they encounter novel words via incidental reading experience. Armed with this, we now turn to discuss the factors that influenced word learning in our experiment. Two factors were investigated: children's level of reading comprehension skill and whether encounters with the word were in semantically diverse or non-diverse contexts.

Reading comprehension skill was strongly associated with word learning throughout the experiment. It was associated with performance on all three offline post-tests, and across a number of reading measures: children with better comprehension skill showed shorter reading times in a number of measures, and a greater reduction in reading times from pre-to post-exposure than those with poorer comprehension skill. While we already know that children with poor comprehension skills have difficulty inferring new words from context (e.g. Cain et al., 2004), our results make a novel contribution to the literature in two key ways.

First, we did not select children on the basis of their comprehension skills and categorise them into 'good' and 'poor' comprehenders, differing in reading comprehension but matched for word reading ability. This dichotomy can be misleading and is certainly not representative of the range of skills we see in a mainstream classroom. Instead we entered comprehension skill as a continuous variable into our models to examine its effect across a wider range of children. We see a clear and compelling relationship between reading comprehension skill and incidental word learning in our heterogeneous sample, and hence we can extend the critical role of comprehension skill in novel word learning to children of this age.

Second, the use of eye movement methodology allowed us to measure the process of learning itself and for the first time, and we have shown that comprehension skill is associated with the process of incidental word learning, not just the end product. After just a few exposures to the novel words, children with stronger reading comprehension skill were distinct from those with less good reading comprehension. The less skilled comprehenders spent longer reading the words in later exposures, meaning that they had more opportunity to benefit from additional time processing the

words, yet this did not help them in their endpoint learning: they performed less well on all three post-tests, tapping both orthographic and semantic learning (see Elgort et al. (in press) for similar findings with second language learners). We suggest that the longer reading times in the later exposures reflect continuing difficulties with encoding the novel words and building a representation of their meaning for those children with lower levels of reading comprehension skill. Previous studies have not been able to ascertain why poorer comprehenders find novel word learning more difficult. Our data show that they are spending time trying to encode novel words, but despite this additional effort, word learning is less successful.

We turn now to address whether word learning was influenced by semantic diversity. Our findings are clear in that we found no evidence of differences in word learning as a function of semantic diversity in the three pen-and-paper post-tests. Similarly, throughout exposure and in the post-exposure phase, there was no main effect of semantic diversity on any eye movement measure during online reading. These findings contrast with those seen in the adult literature (e.g., Johns et al., 2015) and are inconsistent with predictions from the lexical legacy hypothesis (Nation, 2017) which argues that contextual experience in a word's lexical history leads to differences in lexical quality emerging as a consequence of learning.

Why might this be? Clearly, a word needs to be sufficiently frequent in order for effects of semantic diversity to emerge from those exposures. It might be the case that children require more than ten exposures in order to show effects of semantic diversity in a word learning experiment like ours, and that the number of exposures needed may differ as a function of comprehension skill, diversity and context redundancy. This would fit with the pattern of interactions observed between diversity and

comprehension skill, diversity and contextual informativeness, and between diversity, exposure and comprehension. Our first set of interactions showed that an effect of semantic diversity manifested differently depending on the comprehension skill of the child. Put simply, reading times were equivalent for better and poorer comprehenders in the diverse condition but better comprehenders spent less time reading in the nondiverse condition. This suggests that children with good reading comprehension skills spent additional time reading words in diverse contexts during exposure, sensitive to the need to spend more time trying to understand sentences and inferring word meanings when the task was more challenging; this extra processing effort was not needed in the non-diverse condition. In contrast, children with poorer reading comprehension did not adapt their reading in this way, spending a relatively long time reading the target words in both diverse and nondiverse contexts (bearing in mind that it was not the same children reading both context types). Relatedly, the interaction between diversity, exposure and comprehension skill in go past times during exposure showed that only those with good comprehension reduced reading times in the diverse condition, again suggesting that while better comprehenders are able to cope with changing semantic contexts and still learn something, poorer comprehenders show no benefit of increased exposures in this more difficult condition. Comprehension skill appears to play a role when deciding whether it is beneficial to present words in similar versus different contexts.

We also saw an interaction between semantic diversity and informativeness of context: in the diverse condition, children made more regressions out of the novel words if they were embedded in a neutral rather than an informative context, but this was not the case for words encountered in diverse contexts. Although we should be

somewhat cautious interpreting this result because the data include both pre-exposure (before any diverse versus nondiverse manipulations) as well as post-exposure reading behaviour (where we might expect to see a difference between our diversity groups) regression rates, this suggests that children responded differently to contextual informativeness as a function of semantic diversity because they were not equivalent in difficulty. It is plausible that the nondiverse contexts were more challenging and hence triggered more regressions in the neutral (also more difficult) condition. However, we think it is more likely that the diverse contexts proved so challenging that children were overloaded with information. Hence, they were less inclined to regress and re-read in this more demanding condition. Although we see no direct effect of diversity on incidental word learning during reading, it is clear that semantic diversity influenced how demanding children found the sentences to read and this was reflected both in their reading behaviour and its relationship with their overall reading comprehension ability.

When reflecting on differences between our findings and those of previous studies, it is important to remember that our methodology was very different to that employed in other studies. Eye movements when reading meaningful text provide a measure of processing that captures both form and meaning, unlike tasks that consider form and meaning more separately, such as lexical decision and semantic judgement task used in previous studies (Johns et al., 2015; Hoffman & Woollams, 2015). It is possible that effects were masked by the simultaneous conflation of form and meaning processing in our experiment, although if this were the case we would expect to see semantic diversity influencing performance in the offline post-tests. We encourage more research using eye movement methodology as this provides the most direct measure of

how emerging word knowledge influences reading as it happens. Given children of this age learn the majority of new words via reading experience (Nagy & Anderson, 1984), we know remarkably little about how this happens.

In summary, the present study revealed the process through which children learn novel words incidentally through text reading. Children were able to learn something about both the orthographic form and the semantic meaning of novel words presented to them in sentences, using contextual informativeness to guide their reading behaviour and showing a substantial reduction in reading times over the course of learning. Although we did not find any evidence that the semantic diversity of exposures influenced incidental word learning directly, our findings suggest that the relationship between diversity, contextual informativeness and comprehension skill is complex. This question should be pursued in future research by varying the nature, number and time course of exposures and the difficulty of the novel words to be learned.

Acknowledgements

Portions of these data were presented at the 2015 meeting of Society for the Scientific Studies of Reading in Hawaii, USA. We thank Tom Smejka for research assistance; we also thank Megan Bird and Kelly van Earde Layton-Smith for helping with the materials. The experiment was designed and run with the support of awards from the Experimental Psychology Society and the British Academy to Holly Joseph, and with assistance from ESRC grant ES/M009998/1.

References

- Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814-823.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.
- Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.
- Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal, 83*(3), 177-181.
- Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes, 45*(2), 122-159.
- Cain, K., Oakhill, J. V., & Elbro, C. (2003). The ability to learn new word meanings from context by school-age children with and without language comprehension difficulties. *Journal of Child Language, 30*(03), 681-694.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, 96(4), 671.

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: a study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 225.

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (in press). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*.

Hoffman, P., & Woollams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of Experimental Psychology: Human Perception and Performance*, 41(2), 385.

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, 2, 17.

Hsiao, Y. & Nation, K. (2106). Effects of semantic diversity on word recognition in developing readers. Poster presented at the 23rd Meeting of the *Society for the Scientific Study of Reading*, Porto, Portugal, 13th-16th July 2016.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59, 434–446.

Johns, B. T., Dye, M., & Jones, M. N. (2015). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, 1-7.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, 66(2), 115.

Joseph, H. S., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition*, 133(1), 238-248.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, 57(2), 151-162.

Lowell, R., & Morris, R. K. (2014). Word length effects on novel words: Evidence from eye movements. *Attention, Perception, & Psychophysics*, 76(1), 179-189.

Mol, S. E., & Bus, A. G. (2011). To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, 137(2), 267.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, 304-330.

Nation, K. (2017). Nurturing a lexical legacy: reading experience is critical for the development of word reading skill. *Science of Learning*, 2(1), 3.

Nation, K., Angell, P., & Castles, A. (2007). Orthographic learning via self-teaching in children learning to read English: Effects of exposure, durability, and context. *Journal of Experimental Child Psychology*, 96(1), 71-84.

Nation, K. & Castles, A. (in press). Putting the learning in to orthographic learning. In K. Cain, D. Compton & R. Parrila (Eds.) *Theories of reading development*. John Benjamins Publishing.

Nation, K., & Snowling, M. J. (1998). Individual differences in contextual facilitation: evidence from dyslexia and poor reading comprehension. *Child Development, 69*, 996-1011.

Nation, K., Snowling, M. J., & Clarke, P. (2007). Dissecting the relationship between language skills and learning to read: Semantic and phonological contributions to new vocabulary learning in children with poor reading comprehension. *Advances in Speech Language Pathology, 9*(2), 131-139.

Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word identification times in young readers. *Journal of Experimental Child Psychology, 116*(1), 37-44.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357-383.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 275.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366

Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E... & Hulme, C. (2009). *YARC York Assessment of Reading for Comprehension Passage Reading*. London, England: GL Publishers.

Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). *Test of Word Reading Efficiency (TOWRE)*. Austin, TX: Pro-Ed.

von der Malsburg, T. & Angele, B. (in press). False Positives and Other Statistical Errors in Standard Analyses of Eye Movements in Reading. *Journal of Memory and Language*, 94, 119-133.

Vorstius, C., Radach, R., Mayer, M. B., & Lonigan, C. J. (2013). Monitoring local comprehension monitoring in sentence reading. *School Psychology Review*, 42(2), 191-206.

Appendix: Experimental sentences seen during exposure trials for all six target words in diverse and non-diverse conditions

Non-diverse context (military)		Diverse context
1	The British and Americans amalgamated their designs for a new fighter plane.	The British and Americans amalgamated their designs for a new fighter plane.
2	Due to Government cut backs the two regiments amalgamated with each other.	The two football clubs amalgamated and formed a new club with many talented players.
3	The army amalgamated plans with the local charity in order to provide medical care.	The two Scottish clans amalgamated into one so that they would both have the same king.
4	If the two countries stopped fighting and amalgamated into one nation then the war would end.	The government amalgamated the health and education departments to reduce the number of jobs.
5	The goggles and helmets were amalgamated so that there was less kit to carry.	Many religions have amalgamated well because they share the same teachings.
6	Nuclear submarines are powered by amalgamated heat and water so they don't need to refuel.	The director amalgamated a traditional play with pop music to create a brand new musical.
7	The Generals amalgamated their knowledge of warfare to help the Queen beat the enemy.	The two universities amalgamated last year and now there are 15,000 students in total.
8	Army scientists have amalgamated many materials to make bomb proof jackets.	When the Spanish invaded Mexico, the two cultures amalgamated quite well.
9	The generals amalgamated their ideas for removing their equipment from Afghanistan.	The scientific findings were amalgamated in order to find a cure for the disease.
10	The navy amalgamated designs for airports and ships to create aircraft carriers.	The two companies amalgamated and hoped that they would make twice as much money.

Non-diverse context (law/evidence)		Diverse context
1	Enough proof had accumulated so that the jury could make a fair judgement on the case.	Enough proof had accumulated so that the jury could make a fair judgement on the case.
2	The police accumulated a lot of strong evidence which meant they could arrest the thief.	The woman forgot to clean under the bed, so dust had accumulated on the floorboards.
3	Members of MI5 accumulated all the incoming data and saved it onto a computer file.	The girl loved collecting rubbers and accumulated more each week using her pocket money.

4	After the news report went out, the police accumulated more than 25 witnesses.	After just one week at his new school, the boy had already accumulated several new friends.
5	The lawyer accumulated witness statements to get support for the case.	The doctors accumulated enough test results to diagnose and treat the patient.
6	The burglar accumulated information about the neighbourhood before committing the crime.	Lava had accumulated beneath the surface which caused a spectacular eruption from the volcano.
7	The evidence accumulated until there was no question that he was guilty.	His debts accumulated until he had to sell his house to pay off the loan.
8	The proof that she had stolen the money accumulated over time and eventually she lost her job.	Although she had accumulated a lot of wealth, this meant she also had to pay a lot of tax.
9	The witness statements accumulated and in the end he decided to plead guilty.	She was shocked to discover how many emails had accumulated while she was away.
10	The solicitor accumulated the documents for the case and took them to court.	The fluid had accumulated in his lungs and he found it very hard to breathe.

	Non-diverse context (politics)	Diverse context
1	The President intervened swiftly in the civil war and most of the people were grateful.	The President intervened swiftly in the civil war and most people were grateful.
2	The government intervened during the teachers' strike and gave them more money.	He knew that if he intervened before she had finished speaking she would be really cross.
3	They discussed whether it would be acceptable if the US intervened in the Syrian crisis.	The Paramedics intervened immediately when they saw that a man might be having a heart attack.
4	The UK has intervened in other countries' actions when there are human rights abuses.	Schools have often intervened early when a child has problems arriving on time.
5	The government intervened when the economy collapsed to make sure the banks could function.	The shopkeeper intervened as the discussion between the three women was becoming heated.
6	The transport secretary intervened to say that all pensioners should get free travel.	Social Services intervened to help the young people have their meetings in the community hall.
7	The president intervened before the execution and saved the prisoner's life.	The farmer had intervened months earlier so that all the locals could make use of the track.

8	The health secretary intervened and approved the new cancer treatment.	One brave girl intervened and managed to prevent the playground fight becoming more serious.
9	The Council should have intervened much earlier to stop the riots, but they didn't.	Sally hadn't intervened when she witnessed the bullying but she did tell a teacher.
10	The government intervened and as a result, prevented the bill from being passed.	The mother intervened in the early stages of her children's fights, for the sake of peace.

	Non-diverse context (health)	Diverse context
1	Cigarette smoking exacerbated the man's breathing difficulties, but he just couldn't stop.	Cigarette smoking exacerbated the man's breathing difficulties, but he just couldn't stop.
2	His infection was exacerbated because the hospital didn't have any antibiotics to give him.	The wind exacerbated my mad hair-style, and I looked ridiculous when I got to school.
3	The heat had exacerbated the swelling in her broken arm, so the plaster became too tight.	Being told off by Mrs Cooke exacerbated Mary's dislike of the strict new teacher.
4	The pain in her leg exacerbated her bad mood, and she ended up taking more pain killers.	The fire was nearly out, but then the wind exacerbated the flames and it started off again.
5	Doctors not washing their hands properly may have exacerbated the spread of the disease.	Supermarkets spray food smells in shops so that our hunger is exacerbated and we buy more food.
6	A study showed that not using waterproof sun cream exacerbated levels of sunburn.	Watching the film 'Madagascar' exacerbated Louise's longing to go on holiday somewhere hot.
7	Eating greasy food exacerbated his weight problems, which caused him a great deal of upset.	Growing up in a busy city exacerbated Jake's hatred of noise so he moved to the country.
8	Sitting in a dusty room exacerbated her daughter's asthma, so they asked to sit outside.	She tried to calm him down but she just exacerbated the situation and he became very angry.
9	Eating cake exacerbated Sally's diabetes, so next time she'll go for a healthier option.	The horrible sound of an aeroplane overhead was exacerbated by my dad playing loud music.
10	My granny said that going to bed with wet hair had exacerbated my cough and made me even more ill.	The death of Melissa's dog exacerbated her unhappiness, so her mother bought her a new puppy.

	Non-diverse context (school)	Diverse context
--	-------------------------------------	------------------------

1	The Year 6 football team confabulated about their past victory on the way to the match.	The Year 6 football team confabulated about their last year's victory on the way to the match.
2	The school children confabulated in a made-up language so no-one else could understand them.	The witness admitted that she had confabulated with the taxi driver and he had told her the story.
3	Jessica loved the way that she and her mates confabulated with children in school.	Phoebe confabulated with her mother on Skype every day when she first left home.
4	If children confabulated in class and the teacher heard them, she got very cross.	My mum confabulated enthusiastically with her friend while I waited for a lift to football.
5	The head teacher confabulated with the other teachers whilst eating his lunch.	Parents confabulated with one another while they watched their children having swimming lessons.
6	The teachers were all very friendly and often confabulated with the children after school.	He confabulated in such a loud manner, that I had to move to the back of the room.
7	The new girl confabulated cheerfully with everyone, and soon made lots of friends at school.	I wish I had confabulated more with my granny when I was young: she had some amazing stories.
8	The mums confabulated for so long in the playground that the teacher asked them to leave.	The children met on holiday and confabulated happily although they spoke only a little English.
9	The dinner ladies confabulated cheerily with the children as they served them their meals.	After the match, the hockey team confabulated about what had gone well and what had gone badly.
10	The students confabulated about why Johnny had been sent to the head teacher's office.	The journalist was exhausted as the famous actor confabulated for four hours without a break.

	Non-diverse context (animals)	Diverse context
1	The animals had languished because they were weren't being looked after properly.	The animals had languished because they were weren't being looked after properly.
2	The dog languished in her basket for a week, before her owner realised she was pregnant.	The hostage languished for so long, his family hardly recognised him when he was released.
3	The farmer claimed that the horse languished after her foal was taken away from her.	Due to the drought the crops had languished, now there would not be enough food for everybody.
4	When the old lady became unwell, her canary languished as there was nobody to feed him.	Molly languished at home when she heard that her son had been killed in the war.
5	The battery hens on the farm languished as their living conditions were very poor.	May was so hot, the flowers languished until the fire brigade arrived with huge tanks of water.

6	My cat languished for days, so I gave her a new brand of cat food and she perked up.	The child languished in hospital until the doctors changed her medicine which helped a lot.
7	The firefighters think the animals languished because it was difficult for them to breathe.	The soldiers languished for so long, that they were not fit to fight when the battle started.
8	Protesters released the pigs, as they had languished in filthy conditions for too long.	I'm an awful gardener: my cabbages languished for a week before I realised they needed watering.
9	The hamster had languished for several days, and the vet recommended a vitamin injection.	The prisoners languished for days on end in their cells before receiving medical attention.
10	The crocodile languished sadly by the river, mourning her baby who had died.	The plants had languished and eventually died, because the sprinkler system failed.
