Learning to Remember:

The Effect of Time and Vocabulary Size on the Specificity of Novel Words

By

Erica H. Wojcik

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2015

Date of final oral examination: 12/18/14

This dissertation is approved by the following members of the Final Oral Committee:

Jenny R. Saffran, Professor, Psychology
Mark S. Seidenberg, Professor, Psychology
Timothy T. Rogers, Professor, Psychology
Vanessa Simmering, Assistant Professor, Psychology
Haley A. Vlach, Assistant Professor, Educational Psychology

Table of Contents

Acknowledgements

There are many people without whom this dissertation would not have been possible. I would first like to thank my primary advisor, Jenny Saffran. From the beginning of my graduate career, she has encouraged me to explore new methods and areas of research. I am constantly amazed by her wisdom and kindness, which have made my time at UW-Madison not only intellectually challenging, but also incredibly fun and fulfilling. I am grateful to have had a chance to work with and learn from her. I would also like to thank my co-advisor, Mark Seidenberg, and the other members of my committee, Haley Vlach, Vanessa Simmering, and Tim Rogers, for their mentorship on this project and others.

Thank you to the members of the Infant Learning Lab who have helped with various aspects of this dissertation. Erin Long and Courtney Thom recruited participants and assisted in collecting data from those (sometimes terrible) two-year-olds. Rachel Raczynski helped create the stimuli. Hilary Stein, Erika Steinbauer, & Ellen Breen carefully coded a daunting number of videos. Thank you for all your time and patience. Thank you to Tianlin Wang, Christine Potter, Brianna McMillan, Eileen Haebig, Viridiana Benitez, Lynn Perry, and the other graduate students & post-docs in the lab for their thoughtful conversations and words of encouragement. Additionally, I am indebted to the participating families who volunteered their time.

Perhaps most importantly, thank you to my family and friends for their love and support. Thank you to my parents, Michael and Jane, and my sister, Monica, for celebrating my achievements and encouraging me during more difficult times. Thank you to my 1105 roommates, Mia, Katlyn, Sara, and Jonny, for always making time for kitchen conversations and tea; and to Marlo & Issa for reminding me that children are people, too. And lastly, thank you to Tom Yoshikami for everything.

Abstract

Using new words flexibly and accurately beyond the learning moment is critical for effective communication. However, while toddlers' novel word representations are appropriately flexible across some dimensions (such as exemplar color), they are notably inflexible across others (such as background context color). To better understand the specificity of newly learned words, the current studies used a looking paradigm to directly compare exemplar and context generalization, both immediately after learning and after a delay. Two-year-olds were tested on novel word exemplar or context generalization (between subjects), either one minute or one week after learning, and the effect of vocabulary size was investigated. Contrary to previous research, the looking paradigm revealed equally successful novel noun generalization to both new exemplars and new contexts after a one-minute delay (Experiment 1), with no effect of vocabulary size. However, when toddlers were tested after a one-week delay (Experiment 2), a different pattern emerged. While low vocabulary toddlers performed equally well on both types of generalization (as in Experiment 1), high vocabulary toddlers were more accurate in generalizing the novel words to new background contexts than to newly colored exemplars. Accuracy in exemplar generalization for this group did not reach above-chance levels. These results suggest that what toddlers remember about a word is affected by how many words they know. As children learn more words, they begin to efficiently focus on remembering the referent, not the background context. This set of studies provides the first piece of evidence that toddlers do not just learn to learn; they also learn to remember.

Introduction

What does it mean to learn a word? If a child hears the word "cup" for the first time at home while her mother is holding a blue sippy cup, the child may encode the association between that label and the blue cup. Learning this one association is an important step to learning what "cup" means, but the child must move beyond that specific learning moment. She must be able to correctly generalize the word "cup" to other cups in other contexts. The ability to flexibly understand and use a word in novel situations is a crucial aspect of word learning (see Colunga & Smith, 2005).

Additionally, young word learners must be able to understand and use words in this flexible manner across time; they must remember and generalize newly learned words correctly for days, weeks, and months beyond that initial learning moment. While researchers have long studied what children learn about a specific label-referent association within a five- to ten-minute laboratory session, we know very little about how young children remember and generalize newly learned words across longer time scales. Because using words flexibly and accurately beyond the learning moment is critical for effective communication, we must better understand this process in early word learners. Do children easily use new words flexibly, even after longer delays, or do memory processes led to differences in word quality? As children learn more about how word meanings are structured, does their ability to efficiently remember and generalize new words improve?

The current set of studies explores the flexibility of newly learned words both immediately after learning and across a longer time scale. The first experiment uses a looking preference task to investigate how children generalize a novel word's referent to both new exemplars and contexts shortly after learning. Then, a second experiment tests both types of

generalization after a week delay. Additionally, both experiments examine the effect of vocabulary size in order to understand how novel word specificity changes across development. By using a looking preference paradigm to test the generalization of novel words across multiple dimensions, time periods, and vocabulary sizes, this set of studies sheds new light on the quality of early semantic representations, and how word learning changes across development.

Generalizing Novel Words

The ability to generalize words appropriately is an essential characteristic of the lexicon; adult word knowledge is specific—with many important episodic details—as well as abstract and generalizable to new situations (McClelland, 2013; McClelland, McNaughton, & O'Reilly, 1995). We know many of the behavioral quirks of our childhood dog, but we can also easily recognize a new breed of dog that we have never seen before. Because of the importance of properly applying words properly to new referent exemplars and contexts, many researchers have examined children's ability to generalize novel words. Two types of novel word generalization that have been studied are exemplar generalization and context generalization.

Exemplar Generalization. Young children need to correctly generalize nouns to novel instances of the referent. For example, if a child learns the word "*dog*" while seeing her own dog, she has to be able to apply that same label to the neighbor's dog that may differ in color, texture, and size. This generalization behavior, in addition to being necessary for communication, also reveals whether the category of referents to which the young child has attached the label is properly constrained. If a child overgeneralizes the word "dog" to her pet rabbit, it is clear that her representation of what that word means is too broad. Exemplar generalization tasks probe beyond the association between a label and a specific referent, asking whether children have formed a useful and flexible semantic representation, in line with the properties of word

categories in their language.

One strategy that young children use for exemplar generalization is to extend new count noun labels to objects of the same shape. This strategy is called the shape bias. The shape bias is a nuanced behavior that depends on the characteristics of a child's vocabulary, the parameters of the experiment, and other conditions (Cimpian & Markman, 2005; Perry & Samuelson, 2011). However, the overall pattern of results shows that by about two years of age, children tend to generalize newly learned count nouns to objects of the same shape, abstracting over other details such as color or material. In addition to generalizing over properties of the exemplar, though, young word learners must also be able to generalize over properties of the surrounding context.

Context Generalization. If a child learns the word *dog* in her own house, she must be able to understand that same word-referent pairing in the park. This type of flexibility is called context generalization. There are many different features of the context around novel words; the social and auditory surroundings, for example, contribute to the contextual landscape. However, the current project focuses on one aspect of context: the visual background of the referent. For example, the visual background when one sees a dog in a house is different from when one sees a dog in a park.

While recognizing referents in different contexts many seem like a trivial task, there is evidence that infants and young children do not easily generalize knowledge beyond its original context. Research on early memory representations demonstrates that young infants do not retrieve memories unless the visual context matches that of the learning moment (see Hartshorn et al., 1998; Learmonth, Lamberth, & Rovee-Collier, 2004). Related findings exist in the word learning literature; younger infants associate labels with all parts of a scene, not just the intended object (Hollich, Golinkoff, & Hirsh-Pasek, 2007).

This difficulty in generalizing representations to novel visual contexts continues into early childhood. Recent studies have shown that toddlers as old as three years of age tend to anchor novel words to the visual context of the learning moment. Specifically, after being taught a novel label that refers to an object presented against a visually distinct background, two-year-olds show comprehension of the word in a forced-choice paradigm only if the object choices are presented on the same background; they do not generalize the newly learned word to a new context (Vlach & Sandhofer, 2011). While three-year-olds perform slightly better, it not until children are about four years old that they perform equally well in the same background and new background conditions. While context variability during learning can help boost performance in the younger age groups (Goldenberg & Sandhofer, 2013), these studies indicate that at the same age that toddlers easily generalize novel words to new exemplars that are different colors and textures (Samuelson & Smith, 1999), they do not easily generalize to contexts that are different colors and textures.

This difference in the developmental trajectory of exemplar and context generalization is particular interesting considering that both involve similar perceptual processes. In both cases, toddlers have to learn to attend to and encode the relevant features of the stimuli in a word learning moment (Gogate & Hollich, 2010). For exemplar generalization, children learn to attend to and encode the shape of a newly learned count noun. As children learn that words such as "cup" or "dog" refer to categories of solid objects that are organized by shape, they form the higher-order, rule-like abstraction that all count nouns refer to things of the same shape (Colunga & Smith, 2005). Children can then use this abstraction to bootstrap future learning; if they encounter a new count noun, they can expect that it refers to a category of objects organized by shape, and can thus extend their word knowledge appropriately (Samuelson & Smith, 1999).

For context generalization, children must similarly learn to attend to the relevant aspects of the scene, which in this case is the fore-grounded referent (Chun, 2000; Haaf, Lundy, & Coldren, 1996; Jones, Pascalis, Eacott, & Herbert, 2011). This process can be seen as very similar to learning the shape bias. Just as children learn over time that count nouns refer to same-shaped objects, they must also learn that nouns refer to an object, not the background, or object + background. Indeed, if toddlers are presented with object referents against multiple backgrounds instead of just one, they are able to generalize that word to other visual contexts at a younger age (Goldenberg & Sandhofer, 2013). By varying the visual background, the consistency of the object in a naming moment is made salient. Thus, this finding suggests that context generalization stems from learning to focus on the relevant information. It is likely then, that children eventually succeed in context generalization because with experience, they have learned a label refers to an object, not the visual scene surrounding the object.

Despite the similarities between exemplar and context generalization, these processes are studied separately. Exemplar generalization is thought to be integral to the word learning processes, but the role of context generalization in word learning is just beginning to be studied in children. While studies indicate that context generalization emerges later than exemplar generalization (4-5 years vs. 2-3 years; Vlach & Sandhofer, 2011; Samuelson & Smith, 1999), it is difficult to make a direct comparison because the methodologies used to study these processes are different.[1] Thus, more work is needed to fully understand both types of generalization.

The Specificity of Novel Word Representations

[1] Some generalization studies have manipulated both exemplar and context (Goldenberg & Sandhofer, 2013; Vlach & Sandhofer, 2011; Werchan & Gómez, 2014). However, in these studies the two manipulations are correlated, making it difficult to tease apart and compare the two types of generalization.

In addition to using slightly different paradigms that are difficult to directly compare, the studies on exemplar and context generalization thus far mainly employ pointing tasks that require children to not only know the correct referent, but also to explicitly reach and point to it. In traditional novel noun generalization tasks, toddlers are first explicitly taught a novel word, e.g., "This is a wug." They are then presented with two or more new objects that vary in shape, material, or color from the original wug. The experimenter prompts generalization with a sentence such as "Can you get me another wug?" While this paradigm reveals that children prefer to extend novel nouns (specifically count nouns that refer to solid objects) to objects of the same shape, it is possible that children would accept other variations on the original exemplar. In fact, in similar categorization paradigms, children often do not hesitate to select a second, different exemplar if prompted with "Can you get me another one?" (Mandler & McDonough, 1996). Thus, traditional exemplar generalization tasks do not reveal which dimensions of a novel referent have been encoded into the semantic representation; they only reveal the most salient dimensions.

Understanding what information is encoded about the referent of a novel word will help move forward the debate about the specificity versus generalizeability of children's word representations. While some researchers argue that early word representations are tied to early perceptual experiences, and thus contain many specific details (Colunga & Smith, 2005; Jones & Smith, 2002; McClelland et al., 2010; Sloutsky, 2010), others believe that words begin as referring to general, abstract concepts (Booth & Waxman, 2002; Cimpian & Erickson, 2012; Cimpian & Markman, 2005; Mandler & McDonough, 1996). Thus, it is still unclear whether toddler's novel semantic representations are specific or abstract.

In order to shed light on this debate, researchers must use other types of tasks to probe how children generalize words. More specifically, we need to use more sensitive paradigms to examine the flexibility of early representations. This type of approach has been successful used to inform a similar debate within the phonological development literature.

Using Looking Tasks to Probe the Quality of Phonological Representations. For many decades, researchers have studied whether early phonological representations are abstract, or whether they are perceptually encoded with very fine detail (for a review, see Thiessen & Yee, 2010). In order to probe the quality of early phonetic representations, researchers have used a looking preference task in which the characteristics of word labels are systematically manipulated. In this task, participants see two images on a screen (for example, a baby and a dog; Swingley & Aslin, 2000). Participants then hear a prerecorded word that labels one of the pictures. This word is either correctly pronounced ("baby") or incorrectly pronounced ("vaby"). Looking behavior is then coded to compare accuracy (the mean proportion of time spent looking at the target after the word is played) on correct pronunciation versus mispronunciation trials.

Using this paradigm, researchers have found that between the ages of 14 and 24 months, infants are less accurate on the mispronunciation trials. This result has been found for the consonants (Swingley & Aslin, 2000, 2002) and vowels (Mani & Plunkett, 2007) of familiar words, as well as for the consonants and vowels of novel words (Ballem & Plunkett, 2005; Mani & Plunkett, 2008). Together, these studies have used the mispronunciation looking preference paradigm to support the theory that early phonological representations are input-specific rather than abstract. Infants are sensitive to changes in the quality of a word label even for words they have recently learned.

While infants' novel word label representations are specific, the body of work reviewed above suggests that novel word referents are abstract and flexible in some ways—children readily extend new words to objects of different colors and materials (Samuelson & Smith, 1999). However, semantic representations appear to be highly specific in other ways—children's comprehension is affected by changes to the visual context of a novel word (Vlach & Sandhofer, 2011). By using a more sensitive task, such as a looking paradigm, researchers may be able to better understand the specificity of novel word referents. In addition to adjusting task demands, though, inserting a time delay between learning and test may also help elucidate the quality of early word representations.

Time and the Quality of Novel Semantic Representations

Several word-learning studies have employed a delay between learning and test, demonstrating that investigating novel word retrieval across longer time scales leads to new insights into early semantic representations (see Wojcik, 2013). Yet research on both exemplar and context generalization has primarily examined how children extend words immediately after learning. We know, though, that children have to use word knowledge over larger time scales. Does the time between the learning and retrieval of novel words affect how children generalize? In other words, does time affect the specificity of novel semantic representations?

One of the first studies to investigate the long-term retention of novel words was conducted by Carey and Bartlett (1978), who found that many three-year-olds are able to remember words for up to 10 weeks after one brief exposure. A few studies since Carey and Bartlett's have also included a delay between training and test. For example, Goodman, McDonough, and Brown (1998) taught 2-year-olds novel words using semantically informative sentences and then tested comprehension after a 24-hour delay. They found that children could

still understand the words after the delay period, demonstrating that 2-year-olds are able to use semantic context to learn and retain new words. Similarly, Woodward, Markman, and Fitzsimmons (1994) found that both 13- and 18-month-olds show comprehension of a novel word that is directly labeled (i.e., "*This is a dax!*") after a 24-hour delay. Other studies have used similar designs to demonstrate that by one to one and a half years of age, infants can retain a newly learned word for at least a day (Baldwin & Markman, 1989; Horst & Samuelson, 2008; Jaswal & Markman, 2003; Markson & Bloom, 1997; Mervis & Bertrand, 1994; Munro, Baker, McGregor, Docking, & Arculi, 2012; Spiegel & Halberda, 2011; Waxman & Booth, 2000).

While a majority of time-delay studies employ a delay of about 24 hours, some studies have examined the retrieval of newly learned words after longer delays. For example, Vlach and Sandhofer (2012) tested novel word retention both one week and one month after learning. They found that under some learning conditions, the three-year-old participants could remember a newly learned word a week, and even a month, after learning. However, their memory greatly deteriorated across time, and in order for a majority of the participants to remember the words for a week, the learning moment needed to provide strong supports for encoding (such as repeating the word and shaking the referent).

Overall, these studies demonstrate that toddlers are able to remember novel words after delay of a day, a week or even a month. However, these findings only demonstrate that toddlers can retrieve newly learned words after a delay. Are these representations affected by the time delay? Are toddler's lexical representations the same a day or a week later, or have they changed? Researchers have also used a delay between training and test to probe the quality of novel word representations.

Changes in Novel Word Representations Over Time. In addition to asking if young

children can remember words across time, researchers have also used time delays to investigate and understand the quality of what is encoded from a learning moment. For example, researchers have used time delays to investigate the strength of novel word representations learned under different conditions. One way that representation strength can be increased during learning is via deeper semantic processing. If participants are made to process the meaning of a stimulus more acutely, they perform better on recall and recognition memory tests (Craik & Tulving, 1975). By presenting causal information about the referent, (Booth, 2009), explicitly labeling the referent (Horst & Samuelson, 2008), or increasing the salience of the label or referent (Vlach & Sandhofer, 2012), researchers have shown that supportive encoding conditions can increase the strength of a novel word representation. In all cases, the effect of encoding conditions was only found after a delay, thus demonstrating the utility of use time delays to inform our understanding of early word learning.

Despite the importance of using time delays to understand early word learning, there are no studies, to the author's knowledge, that have examined novel noun generalization across time. Thus, while we know what encoding conditions lead to more robustly retained semantic representations, we do not know how the specificity of different semantic features changes with time. One recent study (Werchan & Gómez, 2013), examined exemplar generalization after either a 4-hour time delay or a nap, finding better generalization in the no-nap condition. However, this study aimed to explore the role of sleep in generalization, and thus was not designed to compare different time intervals and different types of generalization. Additionally, the study exposed young children to multiple exemplars of the referent, thus testing how children abstracted across these exemplars to generalize. This does not address the question of the

flexibility of toddlers' novel word representation after an initial exposure to one exemplar. The findings, though, suggest that toddlers' representations become more generalizable over time.

Research on memory and retention also suggests that over time, representations may become more generalizable, particularly across contexts. Studies with both humans and other animals have demonstrated that as time goes on and memories are translated from the hippocampus to cortical areas, they are more easily generalized to novel contexts (Winocur, Moscovitch, & Bontempi, 2010). For example, when rats are exposed to electric shocks in one context (as defined by a distinct cage), they show fear responses to that specific context, and not a novel context, immediately after training. However, with time, fear responses become more and more generalized; after a delay of a month, mice show the same fear response to a novel context as they do to the trained context (Wiltgen & Silva, 2007). Interestingly, while a lesion to the dorsal hippocampus immediately after training leads to deficits in contextual fear responses, if the lesion is made after a longer amount of time, it has no effect (Maren, Aharonov, & Fanselow, 1997). These results suggest that as memories are consolidated to cortex, they become less context-specific and more abstract (see McClelland et al., 1995 for a synthesis and model of this perspective for adult humans).

From this literature, one prediction is that after longer delays, younger children will show more robust context generalization. With time, children's semantic representations may become more context-independent. On the other hand, visual context may remain an integral aspect of novel semantic representations over time. Indeed, we know that some mature lexical representations are context-specific; adults are often faster to activate words if they are heard or read in an appropriate (as opposed to novel) context (Adelman, Brown, & Quesada, 2006;

Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Thus, it is unclear whether toddlers' novel word representations remain context specific with time.

There is less literature that is applicable to how time affects exemplar generalization. It is possible that when toddlers learn novel words, they attend to (and thus encode) shape such that they will use this property to generalize regardless of the delay between learning and test. However, it is also possible that shape is only highlighted in the learning moment, and that after a delay, as the representation decays and is consolidated, the saliency of various properties of the new referent (such as color, texture, and shape) become more equivalent, leading to a weaker shape bias. In order to better understand how exemplar generalization fits into word learning, we need to understand how retention delays affect this task.

By examining how toddlers generalize novel words to both new exemplars and new contexts immediately after learning (Experiment 1), and comparing these findings to when toddler are tested after a week-long delay (Experiment 2), this set of studies probes how the specificity of novel semantic representations is affected by time. Do words become more generalizable, or are the details of the learning moment maintained? Are exemplar and context details retained similarly, or is the representation of one or the other more robust?

An additional aim of the current study was to explore the role of vocabulary size in these generalization and memory processes. Previous research has found a relationship between vocabulary size and word learning and word memory abilities (e.g., Bion, Borovsky, & Fernald, 2013; Houston-Price, Caloghiris, & Raviglione, 2010; Lany & Saffran, 2011; Mills, Plunkett, Prat, & Schafer, 2005; Werker, Fennell, Corcoran, & Stager, 2002). Additionally, studies on the shape bias have demonstrated that as children learn more words, they show a more robust shape bias. Several studies have provided evidence that this is due to toddlers "learning to learn" (e.g.,

Samuelson & Smith, 1999). As children learn more words, they learn how referent categories are organized, and thus know how they should generalize new words. The current study thus also examined the effect of vocabulary size on retention, exemplar generalization, and context generalization across delay conditions.

Summary

One crucial aspect of learning a new word is the ability to accurately generalize that word beyond the conditions of the learning moment. Children's lexical and semantic representations must be specific in some ways, and flexible in others, such that they can apply their knowledge correctly to new situations. The literature reviewed above demonstrates that by two to three years of age, toddlers are able to generalize new words to novel exemplars of different colors and textures. However, at this same age, they are not able to generalize new words to novel visual contexts. The current project directly compares 2-year-olds' generalization to novel exemplars and novel contexts using a looking preference paradigm to further explore the differences in these types of generalization abilities at this age. Additionally, toddlers were tested across either a short or long delay, and the effect of vocabulary was be examined, to probe how the quality of newly learned semantic representations are affected by time and vocabulary size.

Overview of the Current Study

The current studies 1) directly compare exemplar and context generalization in toddlers, 2) examine how the quality of newly learned semantic representations are affected by time, and 3) investigate the effect of vocabulary on both of these effects. Toddlers were trained on four novel words that comprised an image of a novel object against a distinct visual background (*Exposure* phase). They were then tested on how well they encoded the words (*Encoding Test* phase). After a delay of one-minute (Experiment 1) or one-week (Experiment 2), participants

were then either tested on exemplar or context generalization (*Generalization Test* phase). Half of the participants saw trials in which the colors of the exemplars were different (Exemplar Condition), and half of the participants saw trials in which the colors of the visual contexts were different (Context Condition) were different. These *generalization trials* were intermixed with simple *retention trials*, in which the visual stimuli were the same as in the Exposure and Encoding Test phases. The test phases all used the Intermodal Preference Procedure (Fernald, Zangl, Portillo, & Marchman, 2008). In this paradigm, two objects are presented side by side, and participants hear a sentence that labels one of the objects. Their looking behavior is coded offline, frame by frame. By examining toddlers' looking behavior across different generalization types, time delays, and vocabulary sizes, the current studies shed light on how novel word representations are affected by language knowledge and time.

Experiment 1

In order to directly compare exemplar and context generalization, Experiment 1 used looking behavior to investigate both types of generalization in toddlers. The prediction was that toddlers would be better at generalizing to new exemplars than to new contexts, but that high vocabulary toddlers would generalize better than low vocabulary toddlers in the context condition. Additionally, performance was predicted to be lower in the generalization trials than the simple retention trials.

Method

Two-year-olds were taught four novel words that referred to novel objects. Each object was presented against a distinct visual background. Then, they were tested on how well they encoded the novel words using the Intermodal Preferential Looking Paradigm (Fernald et al., 2008). After a one-minute break, participants were tested on their generalization of the newly learned words to either novel exemplars or novel contexts (between subjects).

Participants. Participants were 64 healthy, full-term 30- to 34-month-old toddlers (29.9 – 34.5; M=32), recruited through a database maintained by the Waisman Center. Eligible participants came from monolingual English speaking homes in the Madison area and had no history of hearing problems and no pervasive developmental delays. Expressive vocabulary, as measured by parental report (MCDI Short Form: Level II; Fenson et al., 2000) ranged from 36 to 100 words (mean = 82, median = 88). Twenty-six additional participants were excluded due to inattentiveness (see Results section for exclusion criteria).

Materials. The *Exposure* phase stimuli consisted of four novel label-object pairs. The novel object images were each paired with a unique, visually distinct background (see Figure 1). The object images were photographs of 3D stimuli from Vlach, Ankowski, & Sandhofer (2012),

who used these objects in a novel word-learning task with a similar age group. The objects were chosen to have distinct shapes and colors from one another. The four backgrounds were pdfs taken from Goldenberg and Sandhofer (2013) and were chosen to be visually distinct from each other, as well as maximally contrastive from the novel object with which it was paired. As with the objects, these backgrounds had been used previously in word-learning tasks with a similar age group. The object and backgrounds were combined using Adobe Photoshop, resulting in four high-resolution pdfs.

Each image was paired with one of four novel labels that are phonologically legal in English. Novel word labels were chosen from the NOUN database (Horst & Hout, 2014). Each word was two syllables and had a different onset and offset, resulting in four distinct novel labels: *coodle*, *tulver*, *bosa*, and *manu*. Word labels were spoken in various carrier phrases (see Procedure) by an adult female in child-directed speech. Label-object pairs were counterbalanced across participants.

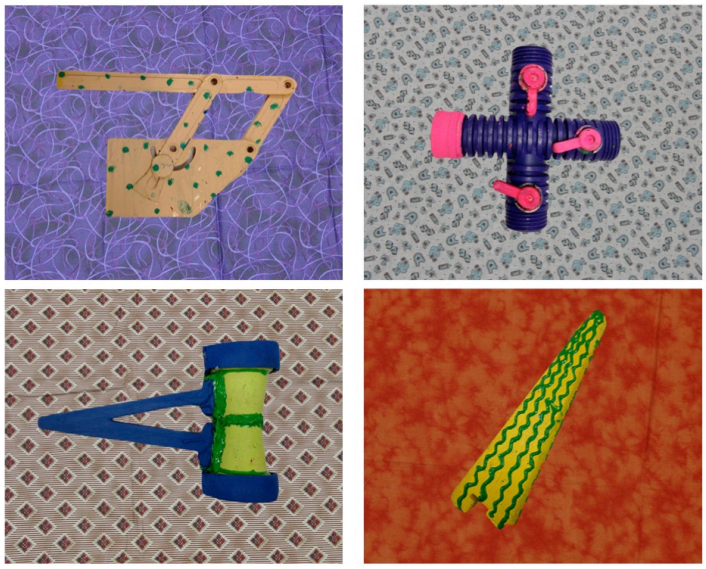

Figure 1. The four novel objects on which participants were trained in the Exposure Phase.

The stimuli for the *Encoding Test* trials were the four label-object pairings from the *Exposure* phase. On each trial, two object images were presented side-by-side, one on the bottom left of the screen and one on the bottom right (see Figure 2). The participants then heard a prerecorded sentence directing them towards one of the objects. The objects were yoked into two pairs, such that each object always appeared with the same distracter (non-target) object.

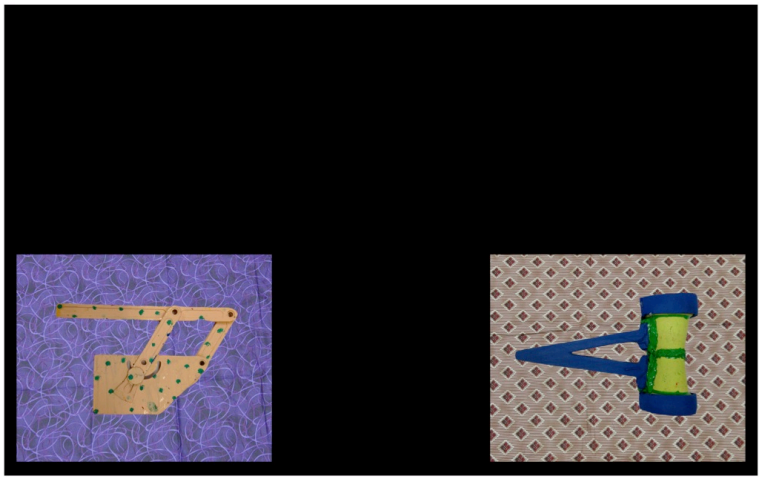

Figure 2. An example Encoding Test trial. After one second of silence, the auditory phase began to play.

The *Generalization Test* materials included one yoked novel word pair that was presented in trials identical to the Encoding Test (*retention trials*). The other yoked pair was tested with different object images (*generalization trials*). For half of the participants (Exemplar Condition), the generalization trials consisted of novel referent exemplars, which were the same shape, but a different color from the exposure images (see Figure 3).[2] Figure 4 shows an example

[2] These new object exemplars were also taken from Vlach, Ankowski, and Sandhofer (2012).

generalization trial with novel referent exemplars. For the other half of the participants (Context Condition), the generalization trials comprised novel context images, in which the exemplar was the same as exposure, but the background was different (see Figure 5).

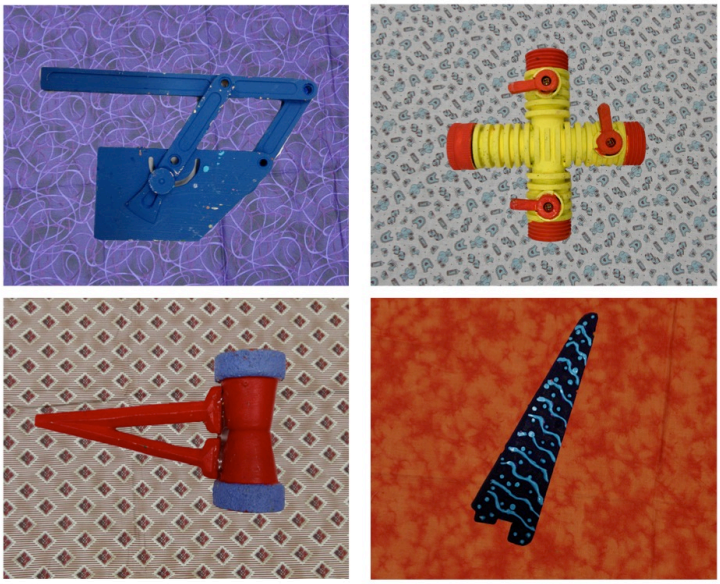

Figure 3. The four novel objects for generalization trials in the Exemplar Condition.

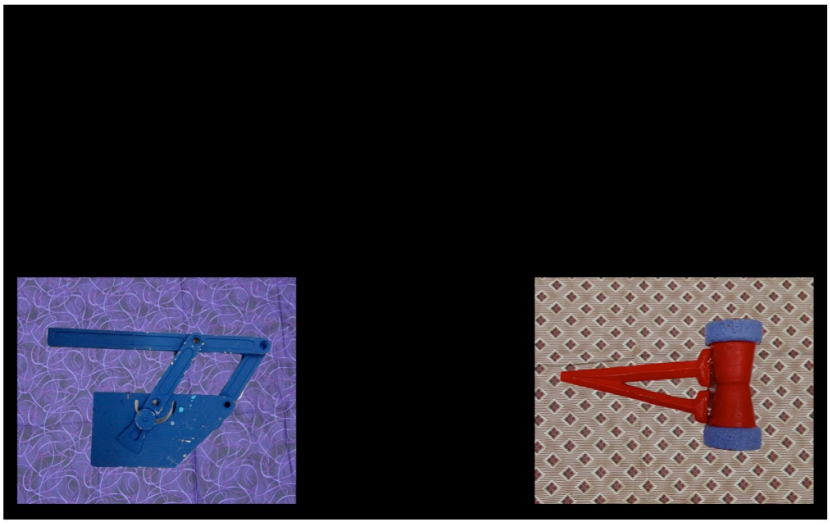

Figure 4. An example generalization trial from the Generalization Test phase for the Exemplar Condition.

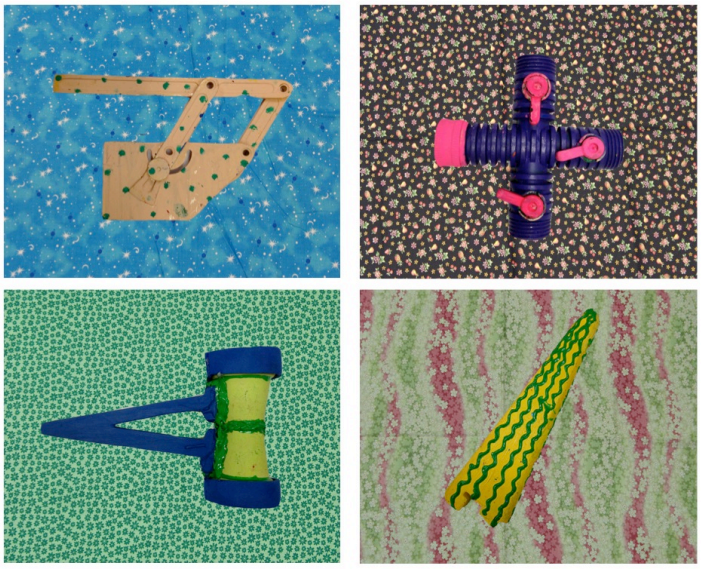

Figure 5. The four novel objects for generalization trials in the Context Condition.

The yoked pair that was assigned to the generalization trials was counterbalanced across participants. The *Generalization Test* materials were designed such that performance on the generalization trials could be directly compared to the retention trials within participants. The comparison of exemplar and context generalization was between participants.

Procedure. Toddlers sat on a caregiver's lap in a sound-attenuated booth, three feet from a 50-inch monitor. Caregivers wore blacked out glasses. The experiment session consisted of an *Exposure* Phase, an *Encoding Test* Phase, and a *Generalization Test* Phase.

Exposure. Participants were first trained on the four novel words. On each trial (6.5s) one of the four objects (Figure 1) was presented on the left or right side of the screen. The object image moved up and down once over the course of the trial to maintain attention. After 1 second of silence, two prerecorded sentences (female speaker, infant-directed speech) were played: "Look at the __! There's a __!" or "See the __! That's a __!" Each trial was followed by a 1 second ISI (black screen). The first two trials labeled familiar objects (a ball and a shoe), in order

to orient the participant to the format of the task. Novel word trials were then presented in four blocks of four, with each object-label pair presented once per block (randomized). A 5s attention-getting video (Baby Einstein clip) was played between blocks, followed by a 1 second ISI. Each object was seen an equal number of times on the left and right side over the course of the Exposure phase. This phase was about 2.5 minutes long, and stimuli were presented with an in-house MatLab program.

Encoding Test. Immediately after Exposure, the Encoding Test (~2.5 min) began. On each test trial (6s), participants viewed two novel object images (the same stimuli as training) presented simultaneous, with one on the bottom left of the screen and one on the bottom right. As mentioned previously, the objects were yoked such that each object was always paired with the same distractor. After 1s of silence, one of two sentence frames was played, directing the participant to one of the objects: "Where's the __?" or "Find the __." The target word onset was always 2 seconds into the trial. A neutral phrase, such as "Can you see it?" or "Look at that!" was then played to maintain attention. This was followed by 1 second of silence.

The test phase began with two familiar word trials (*shoe* and *ball*) to orient participants to the task. The novel word trials were then presented in four blocks of four, with one trial per novel word in each block (trial order was counterbalanced across participants). Blocks were separated by a 5s attention-getting video (Baby Einstein clip or picture + prerecorded phase, such as "You're doing great! Here come some more!"), followed by a 1s ISI. All objects images appeared an equal number of times throughout the test, and the target object was positioned on the left and right an equal number of times as well. After the test trials, the participants watched an unrelated movie for one minute to recapture attention.

Generalization Test. After the 1-minute movie, participants again saw the two familiar

word test trials (*shoe* and *ball*). Then, participants saw four blocks of four test trials in the same format as the Encoding Test Phase, again with 5s attention-getting videos separating the blocks. However, as mentioned in the Materials section, one yoked pair (two of the four novel words) was tested with the *generalization trials*, which consisted of two novel exemplar images or two novel context images (across participants). The trials for the other yoked pair were identical to those in the Encoding Test phase (called *retention trials* in this phase). The pair of words assigned to the generalization trials was counterbalanced across participants. The Generalization Test was 2.5 minutes long.

The entire session in the booth lasted 8 minutes and 20 seconds. Afterwards, the participant's caregiver filled out the Macarthur-Bates CDI (Short Form Level II; Fenson et al., 2000), which included a measure of expressive vocabulary (100-word checklist) as well as a measure of grammatical development (whether toddlers combined words "not yet" "sometimes", or "often").

Results

Data preparation. Looking behavior was coded in 30ms frames by trained coders (see Fernald et al., 2008). Inter-rater reliability was assessed on 20% of the videos (for each condition). The proportion of frames on which the two coders agreed was above 95%. Because of data loss in the second block of test trials in the Generalization Test phase,[3] only the first block (eight trials) was used in the analysis. Mean accuracy scores were then calculated for each

[3] This was an anticipated problem due to the length of the experimental session, and thus the counterbalancing of test trials was designed so that Block 2 could be dropped if necessary. While 77% of trials were usable in Block 1 across all subjects and conditions, only 64% of trials were usable in Block 2. Indeed, for the Exemplar Generalization condition, there was a 20% decrease in the amount of usable trials from Block 1 to Block 2. Additionally, while all subjects in both groups contributed data on at least half of the trials in Block 1, six participants in each condition contributed data on fewer than half of the trials in Block 2. Importantly, several of those participants did not have any usable generalization trials in the second block. To ensure that analysis were not affected by inattention and data loss, Block 2 was not included in the analyses for either test phase in both Experiments 1 and 2.

subject by condition. More specifically, for each trial, the proportion of time spent looking to the target image during the critical window was calculated. This window was pre-defined as 300 ms after the onset of the novel word (in order to adjust for the time it takes to plan an eye movement) to 1800ms after the onset (Fernald et al., 2008). The average of these proportions for each trial type (retention or generalization) was calculated for each subject in both the Exemplar and Context Conditions. This mean was the dependent variable used in all analyses.

Additionally, to examine the role of vocabulary size, participants were divided into low and high vocabulary groups based on a productive vocabulary scores (from parental report; MacArthur-Bastes Communicative Development Inventory: Short Form Level II; Fenson et al., 2000). Participants were assigned to the low or high group based on a median split across participants' scores in both Experiment 1 and 2 (to make comparison across groups and experiments possible). Those with scores at the median (88/100) were assigned to the low vocabulary group.

A median split was chosen over using vocabulary as a continuous independent variable because the distribution of the vocabulary scores was negatively skewed (see Figure 6). This distribution is mostly likely due to the fact that age range of the participants is at the high end of appropriate age range for the chosen measure of productive vocabulary, leading to a ceiling effect. The use of a median split succeeded in separating the children who received a score near the maximum from those distributed across the left tail.

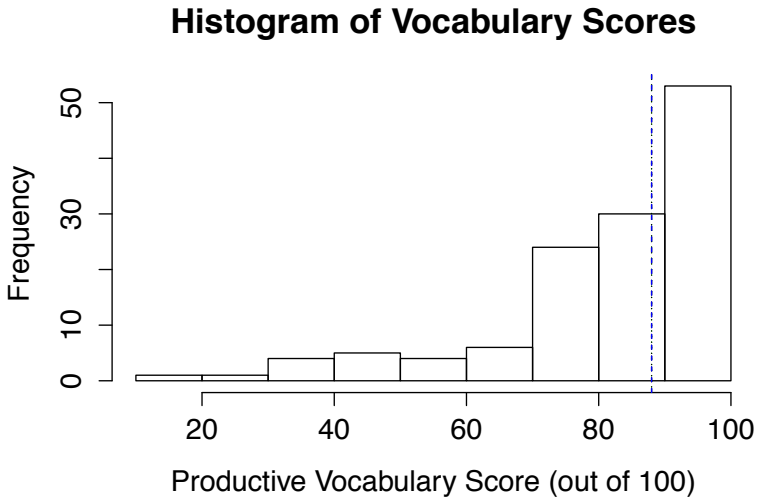

Figure 6. A histogram of productive vocabulary scores. The dotted blue line indicates the median split used for analyses.

Analysis and results. To ensure that participants learned the novel words, performance on Encoding trials was compared to chance (following the analysis procedure in Bion et al., 2013). Although the Encoding Test was identical for all four groups (exemplar and context generalization conditions; low and high vocabulary), each group was analyzed separately to (a) ensure that there were not any encoding differences between participants randomly assigned to the two conditions (Exemplar and Context) and (b) ensure that toddlers with both high and low vocabulary were able to learn the words. For each group, participants' mean accuracy was significantly above chance (see Table 1), demonstrating successful learning. To further confirm that participants in each condition encoded the words equally well, and to assess the effect of vocabulary size on encoding, a 2 (Condition) X 2 (Vocabulary) between-subjects ANOVA was run with mean encoding accuracy as the dependent variable. There was no significant main effect of Condition ($F([1,60]=0.44$, n.s.) or Vocabulary ($F[1,60]=0.076$, n.s.), and the interaction

was not significant (F[1,60]=2.73).

Table 1

Experiment 1 Encoding Test phase results.

Condition	Vocabulary	*Mean*	*SD*	*df*	*t*	*p*
Exemplar	Low	.61	.13	17	3.45	<0.005
	High	.65	.10	13	5.45	<0.001
Context	Low	.68	.11	16	6.60	<0.001
	High	.62	.14	14	3.12	<0.005

Note. Mean, standard deviation, degrees of freedom, t-value, and p-value (one-tailed) for comparisons against chance (0.50). These trials were identical across all groups.

Because training was identical for both groups, the expectation was that there would not be an effect of Condition on encoding accuracy. However, it is notable that there was also no effect of vocabulary, as some studies have found effects of vocabulary size on word learning (Bion et al., 2013; Houston-Price et al., 2010; Lany & Saffran, 2011; Mills et al., 2005; Werker et al., 2002). Other work has found no influence of vocabulary size on toddlers' abilities to learn new words, though (Byers-Heinlein & Werker, 2009; Mather & Plunkett, 2009), so the null effect of vocabulary found in the current study is not unprecedented. In fact, many word learning studies that have found an effect of vocabulary size used either a broader age range (e.g., Mills et al., 2005) or a more difficult task (Bion et al., 2013b; Lany & Saffran, 2011). It is possible that the Exposure phase in the current study was explicit and long enough that all toddlers were able to successfully learn the words. Indeed, because we were interested in retention, we designed the training to give participants enough support to robustly encode the novel words.

With confirmation that participants encoded the novel words in hand, the critical question was how well participants retained and generalized the novel words after the short 1-minute delay. First, performance on both trial types was compared to chance. For the retention trials, which presented participants with the exact referents that they were trained on, all groups except for the low vocabulary, Exemplar condition were at least marginally above chance (see Table 2). For the generalization trials, all groups were at least marginally above chance (see Table 3). In order to compare performance across groups, conditions, and trial types, a 2(Condition: Exemplar vs. Context, between-subjects) x 2(Trial Type: Retention vs. Generalization, within-subjects) x 2 (Vocabulary size: low vs. high, between-subjects) mixed ANOVA was run, with mean accuracy as the dependent variable. There was no main effect of Condition ($F[1,60] = 0.013$, n.s.), Trial Type ($F[1,60] = 0.520$, n.s.) or Vocabulary size ($F[1,60] = 0.097$, n.s.; see Figure 7). Additionally, there were no significant two-way interactions, and the three-way interaction was also not significant.[4]

[4] A similar pattern was found if instead of mean accuracy, I examined the percent of participants who were correct at test in each condition for each trial type, as defined by whether their mean performance was greater than chance (see Appendix).

Table 2

Experiment 1 Generalization Phase results: Retention Trials.

Condition	Vocabulary	*Mean*	*SD*	*df*	*t*	*p*
Exemplar	Low	.53	.22	17	0.58	0.28
	High	.61	.18	13	2.29	<0.05
Context	Low	.59	.25	16	1.50	<0.10
	High	.62	.24	14	1.88	<0.05

Note. Mean, standard deviation, degrees of freedom, t-value, and p-value (one-tailed) for comparisons against chance (0.50). These trials were identical across all groups.

Table 3

Experiment 1 Generalization Phase results: Generalization Trials

Condition	Vocabulary	*Mean*	*SD*	*df*	*t*	*p*
Exemplar	Low	.65	.20	17	3.12	<0.005
	High	.62	.23	13	1.91	<0.05
Context	Low	.60	.19	16	2.07	<0.05
	High	.57	.20	14	1.39	<0.10

Note. Mean, standard deviation, degrees of freedom, t-value, and p-value (one-tailed) for comparisons against chance (0.50).

While no significant effects were found in the omnibus ANOVA, planned follow-up analyses were run for each Trial Type separately for direct comparison with analyses on the Encoding Phrase. Specifically, 2 (Condition) X 2 (Vocabulary) between-subjects ANOVAs were run for both retention and generalization trials. For the retention trials, there was no significant

main effect of Condition (F[1,60] = 0.45, n.s.) or Vocabulary (F[1,60] = 0.85, n.s.). There was also no significant interaction (F[1,60] = 0.22, n.s.; see Figure 7; see Table 2 for descriptive statistics). Recall that on these trials, participants saw the exact same referent stimuli that they were trained on (for two of the four novel words). Because the Retention trials were identical between the Exemplar and Context conditions, the results are as expected. But, as with the Encoding trials, it is interesting that there was no significant effect of vocabulary—participants with low productive vocabularies remembered the novel words just as well as those with high vocabularies.

This pattern of results was confirmed by a Growth Curve Analysis (Mirman, 2014), which included Condition and Vocabulary as fixed effects, but also second-order orthogonal polynomials for each 30-ms frame across the critical time window. This analysis allows for the examination of linear and quadratic trends in looking behavior across the critical time window. For the retention trials, the only significant effect was the overall intercept ($t=2.98$, $p<0.01$), which is equivalent to mean accuracy in the previous analysis. Thus, not only were the retention trials not significantly different in mean accuracy across Condition and Vocabulary, but the looking pattern across the time windows also did not differ across those factors.

For the generalization trials, the 2x2 ANOVA again revealed no significant main effects (Condition: F[1,60]=0.89, n.s.; Vocabulary: F[1,60]=0.26, n.s.) or interaction (F[1,60]=0.003, n.s.). Because these trials addressed the critical question of whether participants would generalize differently to new contexts versus new exemplars, additional planned follow-up tests compared these two types of generalization within each vocabulary group. There was no significant difference in accuracy in the context condition compared to the exemplar condition for either the low vocabulary group (t[33] = 0.75, n.s) or high vocabulary group (t[27] = 0.55,

n.s; see Figure 7; see Table 3 for descriptive statistics). Again, a Growth Curve Analysis reflected this same pattern of results, with no significant affect of Condition or Vocabulary on either linear or quadratic time.

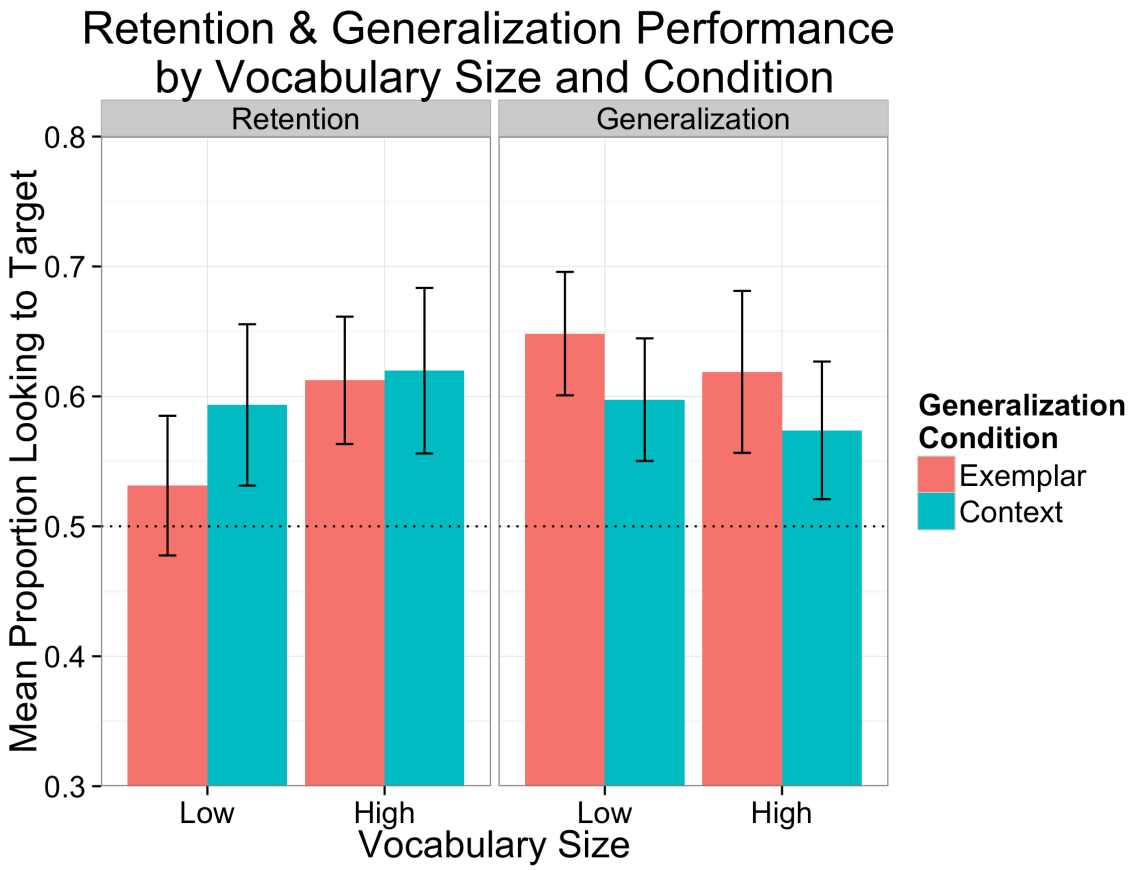

Figure 7. Mean accuracy on Retention and Generalization trials in Experiment 1 (short delay). The dotted line represents change performance. Error bars represent standard error of the mean.

Discussion

In Experiment 1, participants were able to learn four novel words, and they successfully remembered the words across the one-minute delay. It is notable that the Low Vocabulary, Exemplar generalization group did not perform significantly above chance in the retention trials, suggesting they did not successfully retrieve the novel words after the one-minute delay.

However, this group did perform significantly above chance on the generalization trials, providing some evidence that they retained the novel words after the delay. Combined with the fact that the Low Vocabulary group in the Context generalization condition performed at above chance levels on the retention trials (which were identical to the retention trials for the Exemplar group), these results suggest that the poor performance of that specific group may have been due to other factors besides the strength of their novel word representations. Additionally, 33 additional subjects were run in this condition to determine if the below-chance performance is reliably, or simply due to idiosyncrasies in the sample. Using the same median cut off for vocabulary as the first sample, there were eight toddlers with low vocabulary and twenty-five with high vocabulary. In this sample, both the high and low vocabulary groups performed at above chance in the retention trials ($t[24]=4.36$, $p < 0.0005$, $M=0.70$, $SE=0.05$; $t[7] = 1.78$, $p <0.05$, $M=0.58$, $SE=0.04$, respectively), which suggests that toddler with low vocabulary are indeed able to retain words across a short one-minute break. This point will be addressed further in the General Discussion.

The first key finding from Experiment 1, however, is that counter to the hypothesis, there was no difference in performance between the retention and generalization trials, as demonstrated by the non-significant effect of Trial Type (see Figure 7). Toddlers' comprehension of the novel words was not disrupted by a change to either the exemplar or context of the trained referent. Lower accuracy on the generalization trials would have demonstrated that the encoded representations were specific to the color of the exemplar and context of the trained referent. For example, in studies that examine infant comprehension of mispronounced words, (e.g., "vaby" instead of "baby"; Swingley & Aslin, 2000), infants were less accurate on the mispronounced label trials compared to correct pronunciation trials,

indicating that their word-form representations are very well-specified. This is true of novel words forms as well (Swingley, 2007). In contrast, in the current experiment, accuracy on the generalization trials (analogous to the mispronounced-label trials) was *not* lower than on the simple retention trials, demonstrating that toddlers' referent representations are flexible, rather than overly specific.

This difference—that word-form representations are well-specified, but referent representations are flexible—fits not only with previous work on novel noun generalization (e.g., Samuelson, Schutte, & Horst, 2009) but also with how infants hear and use labels and referents in the real world. There are many word forms that differ by only one phonological feature (e.g., "bat" vs. "pat"), and thus children must encode word forms faithfully in order to distinguish between lexical items. Count noun referents (such as those taught in the currents experiments), on the other hand, are often categories of objects that vary along many dimensions, including their color and the contexts in which they are seen. Thus, in order to locate the correct referents across many situations, referents must be flexible across many dimensions. Notably, there are many features of word forms (such as speaker identity) that should be more flexible, as well as many features of count noun referents (such as shape) that should be more specified. This is discussed further in the General Discussion. However, the current results demonstrate that for the dimensions of exemplar color and background context color, toddlers' novel word representations are flexible.

The second key finding from Experiment 1, which also did not support the hypothesis, is that participants performed equally well on both types of generalization: exemplar and context. This result conflicts with previous findings that children at this age, and even older, fail at generalizing novel words to new contexts (Goldenberg & Sandhofer, 2013; Vlach & Sandhofer,

2011). There are three differences between the current study and past studies on novel word context generalization that may account for the contradictory findings.

Firstly, the current study measured context generalization with a less demanding task. Previous studies have used a paradigm in which children are asked to explicitly point to the referent that matches a particular label. This type of task may mask a child's word knowledge. Specifically, children may have some knowledge of the correct referent, but may not be confident enough to point to one of the object choices. Or, they may have implicit knowledge of the correct answer that would allow them to comprehend that word in a more naturalistic language comprehension moment, but not when explicitly asked (see Munakata, 2001 for a more detailed explanation of this argument). The looking task used in the current study addresses these potential confounds. Many toddlers did not make any pointing gestures, and yet participants looked reliably to the correct target in the generalization trials in both conditions. It is possible that young children are indeed able to generalize novel words to new contexts, but that they are less confident or aware of this knowledge, and thus do not perform at above-chance levels in a pointing task.

A second difference from past studies is that the current study tested participants on the trained objects before asking them to generalize to new contexts or exemplars (the Encoding Test phase). Many studies have demonstrated that for infants, children, and adults, retrieving a memory leads to better retention. Retrieval has also been shown to improve adults' and children's generalization (or transfer) of learned information to a new problem (Butler, 2010; Rohrer, Taylor, & Sholar, 2010). While the mechanisms behind this "testing effect" are under debate (Roediger & Butler, 2011), the effect itself is robust. Thus, in the current experiment, the

Encoding Test could have lead to both better retention and generalization compared to other experiments that did not have this additional opportunity to retrieve the novel words.

A third difference is that the current study imposed a one-minute break before the generalization task. Under some theories, one minute is long enough for a memory to be consolidated from short-term to long-term memory (Craik & Lockhart, 1972). It could be this difference is what led to successful generalization. Future studies that remove the short delay in a looking task, and add in the delay in a pointing task, will directly address this explanation. If the one-minute break modulates context generalization in both tasks, this would be strong evidence that transfer to long-term memory explains the results of the current experiment.

While the current results deviate from recent work on toddlers' context generalization of newly learned words (Goldenberg & Sandhofer, 2013; Vlach & Sandhofer, 2011), they are in line with research on infant memory development. Young infants have difficulty retrieving a memory when the context changes, but by around one year of age, this difficulty subsides (in both recognition and recall tests, Barnat, Klein, & Meltzoff, 1996; Hartshorn et al., 1998). Thus, many of the challenges associated with context generalization exist primarily for infants much younger than the participants in the current study, and in other studies of novel word generalization. The memory literature implies that toddlers' early word representation may not be as tightly tied to visual context as was previously believed. By decreasing the task demands, the current study was better able to assess toddlers' generalization abilities, demonstrating that novel words can be generalized to new visual contexts.

However, the results of the current study do not imply that toddlers fail to encode and remember context entirely. Adults still show effects of context in memory retrieval across many tasks (e.g., Boyce, Pollatsek, & Rayner, 1989; Chun, 2000; Godden & Baddeley, 1975).

Toddlers' word comprehension, too, is affected by the congruency of the context to previous experiences (Wojcik, Lew-Williams, & Saffran, 2014). What the current results point to, though, is that by two and a half years of age, toddlers are able to ignore contextual changes if necessary in order to find the referent of a newly learned word.

The last crucial finding from Experiment 1 is that in addition to demonstrating no effect of vocabulary on either encoding or retention (as discussed above), there was also no effect of vocabulary on generalization accuracy. This finding goes against the original prediction, and it is surprising given the role of vocabulary size in some types of exemplar generalization (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). One possible explanation is that sensitivity of the looking measure allowed the low vocabulary group to show better generalization ability than they do in pointing tasks. If this is the case, then it is possible that using a longer time delay will reveal vocabulary-related differences. Thus, following the null vocabulary effect of Experiment 1, the revised prediction is that the higher vocabulary group will show better context and exemplar generalization than the low vocabulary group in Experiment 2 (which imposes a week-long delay before test).

While the results of Experiment 1 demonstrate that toddlers' representations are flexible enough to generalize to novel exemplar colors as well as novel visual contexts, we do not know if this flexibility remains a characteristic of the lexical-semantic representation if there is a longer time delay. Are toddlers able to retrieve novel words after one week? Can they still generalize to novel exemplars and novel contexts, or are their representations tied to specific features of the training stimuli do they need more cues to be able to recall the word referents? Do toddlers with larger vocabularies generalize differently than those with smaller vocabularies? By testing a new

sample of participants with a one-week delay between training and test, Experiment 2 aims to shed light on how novel word representations change over time.

Experiment 2

By inserting a one-week delay instead of one-minute delay between the Encoding and Generalization Test phases, Experiment 2 tested the specificity of novel nouns over time. In Experiment 1, toddlers performed very well on both generalization conditions after the short delay, with accuracy not significantly different from retention trials for either group. I predicted that after a longer delay, participants would perform as well on the generalization trials as on the retention trials, as they did in Experiment 1, perhaps with increased performance in context generalization, as discussed in the Introduction. I also predicted that there would be an interaction with vocabulary size, such that toddlers with higher vocabularies would be more accurate on generalization trials, and in particular on context generalization trials. There was no effect of vocabulary in Experiment 1. However, as discussed in the introduction, the addition of a time delay may reveal effects that do not show up immediately after learning.

Method

A new sample of two-year-olds were taught the same four novel words and tested on encoding as in Experiment 1. However, instead of viewing the Generalization Test one minute later, they were brought back to the lab a week later for this phrase of the experiment.

Participants. Participants were 64 healthy, full-term 30- to 34-month-old toddlers (30.1 –33.9; M=31.7), recruited through a database maintained by the Waisman Center. Eligible participants came from monolingual English speaking homes in the Madison area, and had no history of hearing problems and no pervasive developmental delays. Expressive vocabulary scores, as measured by parental report (MCDI Short Form: Level II; Fenson et al., 2000), ranged from 12 to 100 words (mean = 82, median = 88). Vocabulary scores were not significantly different from the participants in Experiment 1: $t(126) = 0.08$ n.s. Twenty-eight additional

participants were excluded due to inattentiveness (10 during the first lab session and 13 during the second) or inability to return for the second session (5).

Materials. The materials were identical to those used in Experiment 1.

Procedure. The procedure was identical to Experiment 1, except that the first lab session ended after the Encoding Test (just under 5 minutes total). Caregivers then filled out the MCDI (Fenson et al., 2000). Participants were brought back to the lab 3-11 days later (mean and mode=7). This second lab session consisted of just the Generalization Test phase (~2.5 minutes).

Results

As in Experiment 1, participants' eye-movements were coded in 30ms frames. Again, inter-rater reliability was assessed on 20% of the videos (for each condition). The proportion of frames on which the two coders agreed was above 95%. For each participant, a mean accuracy score was calculated for each trial type. Participants were also assigned to the low or high vocabulary group as in Experiment 1, based on a median split of productive vocabulary scores.

To ensure that participants successfully encoded the words, encoding accuracy was compared to chance. Indeed, participants in all groups looked to the correct referent at above-chance levels during the encoding trials, confirming that the novel words were learned (see Table 4). To confirm that there were no differences in encoding between the Exemplar and Context Conditions, and to test if there were any differences in encoding based on vocabulary size, a 2x2 ANOVA was run. There was no significant main effect of Condition ($F[1,60] = 0.16$, n.s.) or Vocabulary Size ($F[1,60] = 0.30$, n.s.). Additionally, the interaction between Condition and Vocabulary was not significant ($F[1,60] = 0.95$, n.s.). Thus, as in Experiment 1, all participants encoded the novel words equally well.

Table 4

Experiment 2 Encoding Phase results.

Condition	Vocabulary	*Mean*	*SD*	*df*	*t*	*p*
Exemplar	Low	.62	.12	15	4.00	<0.001
	High	.60	.15	15	2.65	<0.01
Context	Low	.60	.16	15	2.44	<0.05
	High	.65	.16	15	3.91	<0.001

Note. Mean, standard deviation, degrees of freedom, t-value, and p-value (one-tailed) for comparisons against chance (0.50). These trials were identical across all groups and to the encoding trials in Experiment 1.

As in Experiment 1, the critical question was how well participants retained and generalized the novel words, but in the current experiment the delay was one week instead of one minute. The same analyses were run as in Experiment 1, first comparing performance in all groups to chance. Then, a mixed ANOVA was run to compare performance across Condition, Trial Type, and Vocabulary Size. For the retention trials, which measured whether participants remembered the words across the delay, all groups performed significantly above chance (see Table 5). For the generalization trials, all groups performed at least marginally above chance expect for the high vocabulary, exemplar generalization (see Table 6). The omnibus mixed ANOVA, 2(Condition: Exemplar vs. Context, between-subjects) x 2(Trial Type: Retention vs. Generalization, within-subjects) x 2 (Vocabulary size: low vs. high, between-subjects), revealed no significant main effects. For Condition, $F(1,60)=0.057$, n.s.; for Trial Type, $F(1,60)= 1.72$ n.s;

for Vocabulary, F(1,60) = 0.40, n.s. As in Experiment 1, there were no significant two-way interactions, and the three-way interaction was also not significant (see Figure 8).

Table 5

Experiment 2 Generalization Phase results: Retention Trials.

Condition	Vocabulary	*Mean*	*SD*	*df*	*t*	*p*
Exemplar	Low	.69	.19	15	3.92	<0.001
	High	.73	.17	15	5.24	<0.001
Context	Low	.66	.22	15	2.92	<0.01
	High	.71	.17	15	4.58	<0.001

Note. Mean, standard deviation, degrees of freedom, t-value, and p-value (one-tailed) for comparisons against chance (0.50). These trials were identical across all groups.

Table 6

Experiment 2 Generalization Phase results: Generalization Trials.

Condition	Vocabulary	*Mean*	*SD*	*df*	*t*	*p*
Exemplar	Low	.68	.17	17	4.18	<0.001
	High	.57	.26	13	1.11	0.14
Context	Low	.61	.26	16	1.69	0.05
	High	.72	.20	14	4.69	<0.001

Note. Mean, standard deviation, degrees of freedom, t-value, and p-value (one-tailed) for comparisons against chance (0.50).

As in Experiment 1, to further investigate the retention and generalization trials, separate planned follow-up ANOVAs were run for each, with Condition and Vocabulary Size as between-subject independent variables. For the retention trials, there was no difference in performance across Condition ($F[1,60]=0.32$, n.s.) or Vocabulary Size ($F[1,60] = 0.71$), and no significant interaction ($F[1,60] = 0.002$; see Figure 8 and Table 5). As in Experiment 1, this pattern of results was confirmed with a Growth Curve analysis that also included second-order polynomials as fixed effects to look at looking behavior over time. For the generalization trials, however, a different pattern emerged. While there was no main effect of Condition ($F[1,60] = 0.63$, n.s.) or Vocabulary size ($F[1,60] = 0.011$, n.s.), there was a significant interaction of Condition and Vocabulary Size: $F(1,60)= 3.91$, $p = 0.05$; see Figure 8). Indeed, the only marginally significant effect in the Growth Curve model was the interaction between Condition and Vocabulary size on the intercept ($t=1.74$, $p = 0.07$). Follow-up tests on the ANOVA revealed that for participants with lower vocabularies, there was no significant difference in accuracy between the Context and Exemplar conditions: $t(30) = 0.85$, n.s. For the high vocabulary group, though, there was a marginally significant difference between context and exemplar generalization ($t[30] = 1.93$, $p <0.1$), such that participants were more accurate in the Context generalization condition ($M = 0.73$, $SE = 0.012$) compared to the Exemplar generalization condition ($M= 0.57$, $SE = 0.016$). Indeed, while the high vocabulary group in the context condition had a mean accuracy that was significantly above chance, the high vocabulary group in the exemplar condition did not (see Table 6). Thus, while low vocabulary groups succeeded in both types of generalization after the delay, with accuracy scores that did not significantly differ from each other, only the high vocabulary group in the Context condition succeeded in

generalization. The high vocabulary group in the Exemplar condition was at chance on the generalization trials.[5]

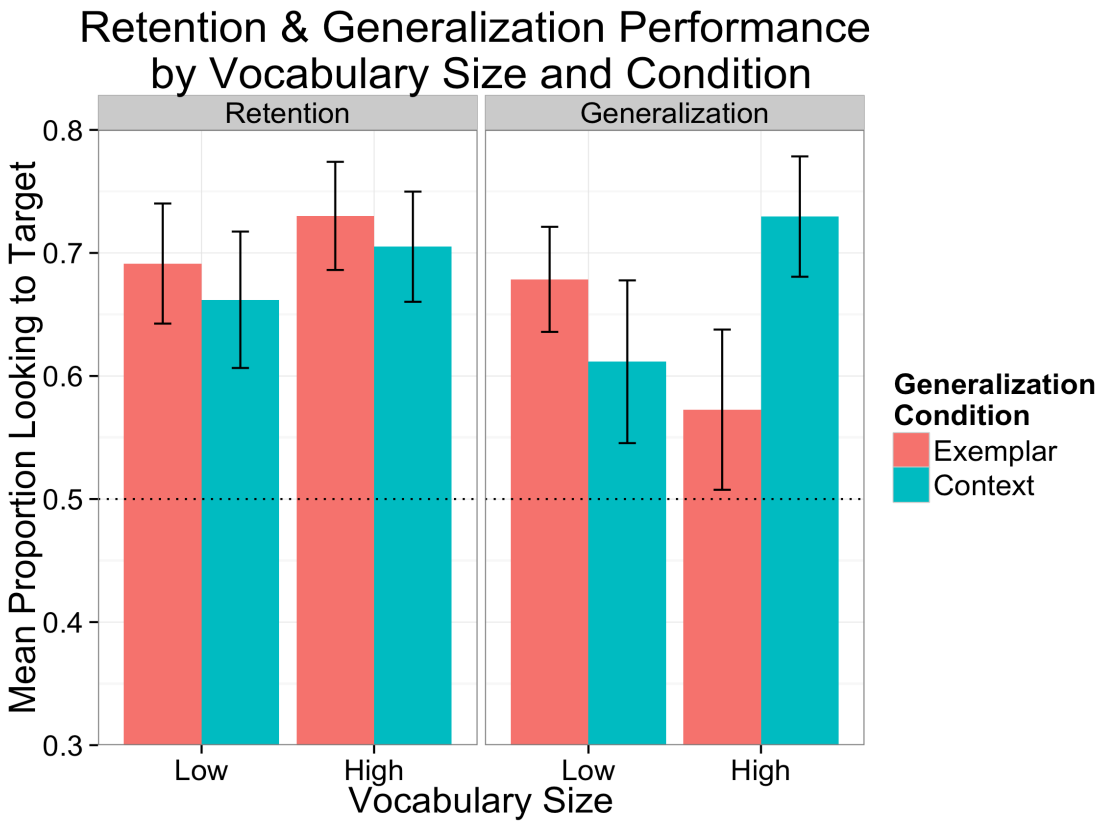

Figure 8. Mean accuracy on the Retention and Generalization trials for Experiment 2 (long delay). The dotted line represents chance performance. Error bars represent standard error of the mean.

[5] As in Experiment 1, a similar pattern of results was found when I examined the percent of participants in each condition who showed above-chance performance across the trials (see Appendix).

The generalization trial results are distinctly different from those in Experiment 1, which found no difference across vocabulary or generalization condition after a one-minute delay. This change in performance suggests that toddlers with higher vocabularies become less flexible in exemplar generalization after a long delay. To examine the effect of time directly, we can compare performance across experiments to examine exactly how context and exemplar generalization are affected by different time delays. Are high vocabulary toddlers tested after a weeklong delay indeed significantly worse at exemplar generalization than those tested after a minute delay?

When follow-up tests were run to compare generalization performance across the short and long delays, no difference was found in mean accuracy between the one-minute delay group and the one-week delay for the low vocabulary groups, either in the Exemplar ($t[32] = 0.47$, n.s.) or Context generalization condition ($t[31] = 0.18$, n.s.). This was also reflected in a Growth Curve model on the low vocabulary groups across Time and Condition, where there were no significant interactions on the intercept or time terms. For the high vocabulary groups, there was also no significant difference in Exemplar generalization between the delay conditions ($t[28]=0.51$, n.s.). However there was a significant difference for the high vocabulary Context generalization conditions ($t[29]=2.16$, $p<0.05$), such that participants' accuracy was significantly higher in the one-week delay group ($M=0.73$, $SE = 0.012$) compared to the one-minute delay group ($M=0.57$, $SE = 0.014$). Again, supporting these results, the only notable effect in the Growth Curve for the high vocabulary group was a marginally significant interaction between Time and Condition at the intercept ($t=1.7$, $p = 0.08$). These analyses suggest that the weeklong delay had an effect on context generalization, but only for participants with higher vocabularies.

Discussion

There are several notable aspects of the results of Experiment 2. Firstly, as in Experiment 1, there was no main effect of Trial Type, demonstrating that toddlers did not perform significantly worse on generalization trials than retention trials. Secondly, performance on the retention trials was high across the board after the one-week delay. This finding is surprising because three-year-olds often fail at retrieving a novel word after this long of a delay (Booth, 2009; Vlach & Sandhofer, 2012). For example, Vlach and Sandhofer (2012) found that less than 50% of participants handed the experimenter the correct referent at test after a one-week delay. In contrast, for each group of participants in Experiment 2—high and low vocabulary; exemplar and context generalization condition—over 75% of participants showed retention after one week by looking more to the correct target on average (see Appendix). The high performance in the current study may be due to the fact that past studies (including Vlach & Sandhofer, 2012) asked toddlers to explicitly point to the correct referent for a label. As noted in the discussion for Experiment 1, by using a looking-time measure, and thus not forcing participants commit to one answer, it is possible that toddlers were better able to demonstrate that they remembered the novel words after a long delay.

Another possible explanation for the high performance on retention trials is that the Encoding Test may have further strengthened the novel word representations. Retrieval greatly strengthens memory representations (for a review, see Roediger & Butler, 2011). Indeed, toddlers in both experiments were asked to retrieve each novel word four times across the encoding test phase. By re-activating the newly learned representations during this phase, participants may have strengthened their representation such that the delay across the time delay was not as strong as found in past studies (e.g., Vlach & Sandhofer, 2012). The effect of retrieval

on word learning, as well as differences between pointing and looking tasks, should be further investigated in light of the current findings.

The most striking result from Experiment 2 is the difference in performance between the high and low vocabulary groups on the generalization trials. After a weeklong delay, toddlers with higher productive vocabularies, but not those with lower vocabularies, generalized more successfully to new contexts than to new exemplars. This result is particularly salient when juxtaposed with the analysis of encoding and retention test trials, which revealed no main effect of vocabulary. Together, these findings suggest that while toddlers with both smaller and larger vocabularies can learn and retain novel words equally well after a delay, their representations of words are different. Specifically, after a weeklong delay, the referent representations of toddlers with high vocabularies are flexible across contexts, but not across exemplar color; participants did not reach above-chance levels in the exemplar generalization condition. Toddlers with lower vocabularies, on the other hand, have referent representations that are equally flexible across both dimensions.

Why did the high vocabulary participants not succeed in exemplar generalization? This goes against previous work demonstrating that children at this age easily generalization to objects with a different color, but the same shape, as the trained referent (e.g., Samuelson et al., 2009; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). It also differs from what we found in Experiment 1, where high vocabulary toddlers generalized well to exemplars of a different color after only a short delay. This result suggests that while the shape bias may be a robust in-the-moment strategy, toddlers' representations may become less flexible over time. This possibility is discussed further in the General Discussion.

While this at-chance performance of the high vocabulary toddlers in the exemplar generalization condition suggests that time affects exemplar specificity, the follow-up analyses comparing Experiments 1 and 2 point to a different story. For the high vocabulary toddlers, a long delay lead to better context generalization, rather than worse exemplar generalization. This result is consistent with the prediction that toddlers would generalize more accurately to novel contexts after a delay. However, one caveat is that there are different attention demands across delay conditions, making it difficult to interpret comparisons across delays. While participants in the first group are performing the generalization task after the Exposure and Encoding Test phases, participants in the second group go right into the Generalization Test phase at their second visit. The effect of this difference can be seen in the number of usable trials in the Generalization Phase—recall that trials were dropped if participants were not looking at one of the referents for 1/3 or more of the generalization window. While 92% of trials across participants were usable in Experiment 2, only 77% of trials were usable in Experiment 1. Due to the varying attentional demands, quantitative differences across delay conditions are harder to interpret than differences within the same delay condition. Participants in the short delay condition could have lower accuracy scores overall due to decreased attention. Thus, the differences in the *pattern* of results across the two experiments, instead of the quantitative differences, are more interpretable.

The pattern of results across the experiments reveals that the high vocabulary group showed more accurate generalization in the context condition than the exemplar condition after the long (but not the short) delay. This notable finding suggests an interesting novel theory: as toddlers learn more words, they may be learning what to remember. This idea is addressed further below.

General Discussion

The current set of experiments had three main aims: (1) to directly compare and contrast the flexibility of novel words across context and exemplar features (Experiment 1), (2) to examine how the specificity of novel words is affected by time (Experiment 2), and (3) test the effect of vocabulary size on both of these effects to investigate if novel word specificity changes as children learn more about characteristics of word meanings.

To address the first aim, toddlers were tested on exemplar and context generalization with a looking preference paradigm (Experiment 1). This is the first study to the author's knowledge to make a direct comparison between these two types of generalization. The results revealed that 2.5-year-olds can generalize novel words to new exemplars and new contexts equally well, and that are just as accurate on these generalization trials as they are on simple retention trials, in which they were tested on the same referent as training. Additionally, there was no effect of toddlers' vocabulary size on performance. Thus, after a short delay, toddlers' novel word representations are flexible across exemplar and context color, regardless of how many words they can produce.

To address the second aim of investigating effect of time on novel word specificity, Experiment 2 implemented a one-week delay between learning and the same generalization tests used in Experiment 1. While toddlers with low vocabularies performed equally well in both generalization conditions, as in Experiment 1, toddlers with high vocabularies did not. High vocabulary toddlers were more accurate if tested on context generalization than if tested on exemplar generalization. In fact, accuracy in generalizing to novel colored exemplars did not reach above-chance levels for this group. In other words, toddlers with larger productive vocabularies had representations that were specific to the trained exemplar, but flexible across

contexts. Comparisons across Experiment 1 and 2 revealed that for high vocabulary groups only, the context generalization/long delay group was more accurate than the context generalization/short delay group. These results speak to the third aim of the study, to investigate the effect of vocabulary size, and suggest that an increase in vocabulary size leads to representations that are more focused on a specific referent exemplar. Throughout both Experiments, Growth Curve Analysis reflected the same pattern of results, with significant effects only on the intercept, and not on either the linear or quadratic time terms.

Taken together, these studies reveal that the specificity of novel word representations is indeed affected by time, but only for toddlers with higher vocabularies. Toddlers with low vocabularies robustly generalized to new exemplars and contexts after a one-minute or one-week delay, demonstrating that their lexical representations are flexible across multiple dimensions after both short and long time spans. Toddlers with high vocabularies, on the other hand, only generalized well across both exemplar and context changes after a one-minute delay. Those tested after a one week delay were able to generalize to new contexts, but not to new exemplars; they had less flexible representations if tested after a week. [6]

Learning to Remember

This striking pattern of results, that toddlers with higher vocabularies have representations that are more specific to a novel word's referent after a long delay, suggests that

---

[6] Although the analyses show an effect of vocabulary size, it is worth noting that in the participant sample, productive vocabulary size was positively correlated with age; $r = .21$, $t(126)= 2.47$, $p<0.05$. Additionally, female participants (n=68) had significantly higher vocabularies than male participants (n=60): $t(126)=2.32$, $p<0.05$; female mean = 85.75, SE= 1.96; male mean = 76.64, SE=2.38. This is not uncommon: females have higher vocabularies than males at this stage in development (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). My hypothesis was that vocabulary would have an effect of generalization, and this was motivated by previous research. Thus, I believe that the effect is driven by vocabulary, not age or sex. Future studies, though, can tease this apart.

as toddlers learn about the structure of word meanings, they learn what to remember. Toddlers with larger vocabularies may realize that to understand words beyond the initial learning moment, they must remember details of the word referent. Details of the background context, on the other hand, are not as important. A dog is a dog, regardless of whether it is at a park or in a house. That only toddlers with higher vocabularies focused their representations on the referent after a long delay suggests that they have become more efficient at remembering and retrieving the relevant aspects of a novel word's meaning.

This learning-to-remember effect is analogous to the learning-to-learn account of the shape bias. As children learn more words that are concrete nouns organized by shape (such as "ball" or "cup"), they learn that new concrete nouns should also be organized by shape. (Perry & Samuelson, 2011; Samuelson & Smith, 1999). In other words, they learn that they can ignore changes in color or texture, because concrete nouns tend to vary across those dimensions. The current findings add to this account, showing that this same "learning-to-learn" mechanisms may also be operating across longer time scales: as children learn more words, they learn to remember more efficiently.

A slightly different interpretation of toddlers' performance in the generalization task is that it reflects not what they remember, but rather the aspects of the stimuli they choose to generalize over in the moment. Toddlers may represent the background context in their lexical representations, but know that during this specific generalization task, they can ignore that dimension of the referent. In fact, there is evidence that infants' familiar word representations do include experience-congruent visual context information (Wojcik, Lew-Williams, & Saffran, 2014). The current study's results do not necessarily show that high vocabulary two-year-olds completely forget information about context over time. Instead, they may suggest that the

flexibility of the representation changes, such that toddlers with high vocabularies are less willing to accept changes to an exemplar, and more welling to accept changes to background context during comprehension. Performance, however, is intimately linked to knowledge. It is near impossible to divorce representations from the processes that underlie their retrieval, such as attention demands or context cues (Smith, Colunga, & Yoshida, 2010). Thus, whether the results are viewed as demonstrating learning to remember or learning to retrieve, it is clear that how children understand and extend words across longer time scales is affected by their vocabulary size.

The interaction between time delay and vocabulary size on generalization ability adds to the debate about whether early word representations are general or abstract (e.g., Booth & Waxman, 2002; Sloutsky, 2010). While this debate has typically focused on what children encode from a learning moment (a generalizable concept or a specific percept), the current studies suggest that we also have to consider the effect of memory processes on the specificity of early words. Instead of simply asking whether words are specific or generalizable across certain dimensions, we must ask how and why the quality of word representations changes over time. The current study's learning-to-remember pattern suggests that for toddlers with higher vocabularies, novel lexical representations become more specific to the referent, but more generalizable across contexts, with time. Why, though, does time have this effect, and why is it specific to the high vocabulary group?

Possible Mechanisms Behind the Learning-to-Remember Effect

There are several possible explanations for the learning-to-remember pattern of results. First, it may be that a longer delay simply allowed differences between the high and low vocabulary groups' initially encoded representations to emerge. It may have taken either some

forgetting or a break from the task to pull out those differences. In the short delay condition, participants were only recently exposed to the new words, and thus any differences in their encoding of context and exemplar details may have been overridden by high overall performance. Indeed, previous studies have found that some differences in novel words representations appear only after a week-long delay (Booth, 2009; for a review, see Wojcik, 2013). Booth (2009) suggested that differences revealed after a delay could be due to the fact that toddlers need to recover from the fatigue of learning before being able to demonstrate new knowledge, and this could be true of the current experiments as well. However, she also notes that improved performance after a delay may be due to consolidation processes (Stickgold, 2013; Walker & Stickgold, 2010).

Consolidation processes underlie the second possible explanation for the fact that differences in generalization were found only in the long delay condition. Consolidation is the process of a memory trace being re-encoded into a more stable representation in the cortex, which can take days or weeks beyond the initial learning moment (for reviews, see McGaugh, 2000; Wixted, 2004). The ability to successfully consolidate a memory significantly improves across the first two years of life and may account for increases in retention ability across that same period (Bauer, 2004, 2005). Thus, the fact two-year-old participants in the current study were given enough time to consolidate the memory traces of the novel words may have had an effect on the quality of those representations. Consolidation leads to representations that are more context-independent, more generalizable, and more integrated into one's semantic network (Davis, Di Betta, Macdonald, & Gaskell, 2009; Gomez, Newman-Smith, Breslin, & Bootzin, 2011; Winocur, Moscovitch, & Sekeres, 2007). In line with this literature, the analysis comparing Experiments 1 and 2 suggests that high vocabulary toddlers generalized better to

novel contexts if there was a longer delay between learning and test. These results could be explained by the fact that representations become more context-independent when are transferred from the hippocampus to the neocortex (Winocur et al., 2007).

However, if consolidation accounts for the results via more context-independence, we would have seen better context generalization for all participants, regardless of vocabulary size. The fact that the effect was only seen for the high vocabulary group suggests that consolidation process may have instead led to more integration with those toddlers' larger semantic networks (Davis et al., 2009), leading to a retrieval pattern that better reflects the characteristics of their lexical-semantic network—specifically, a focus on the referent over the context. Without a large enough vocabulary underlying their semantic network, the low vocabulary group may have not benefited from the integration process. This mechanism fits with the learning-to-remember pattern of results. As children learn more words, novel words can be efficiently integrated into this large network, leading to retrieval behaviors that reflect patterns in that network. Future work will manipulate consolidation directly (via sleep, for example; Walker & Stickgold, 2010), to test this possible mechanism.

A final possible explanation for the current results is targeted forgetting. The theory of retrieval-induced forgetting asserts that memories associated with a cue compete for recall, and thus when one association is retrieved, not only is that association strengthened but also (and perhaps primarily), the competing association is forgotten (Anderson, Bjork, & Bjork, 1994). This forgetting affects the quality of long-term memory representations. A version of this theory has been applied to generalization as well. Vlach (2014) argues that generalization is facilitated if exposure to multiple exemplars is spaced, because this allows the variable, unimportant features to be forgotten. Importantly, in most studies on the role of forgetting in generalization,

participants are presented with multiple exemplars from a category, and so generalization is a function of abstracting over these exemplars (e.g., Vlach, 2014; Vlach & Kalish, 2014). Thus, retrieval-induced forgetting can be directed toward the features that are not consistent across exposures.

In the current study, though, participants were only trained on a single exemplar of each referent; there were not features that were retrieved more often than others. It was not until the generalization test that participants saw exemplars that varied across any dimension. It is possible, though, that forgetting still accounts for the high vocabulary group's performance after a long delay. The high vocabulary group's performance could be interpreted as demonstrating forgetting the color of the context, and strengthening the color of the exemplar. Since participants did not see exemplars with variable context colors and an invariable exemplar color, generalization cannot be attributed to abstracting over a variable feature (as in Vlach & Kalish, 2014). Instead, the time delay itself may account for this targeted forgetting and generalization. High vocabulary participants in the current study may have *learned* to what to forget and what to remember. Because they have learned many words before, for which the referent, not the context, is the invariant aspect of a referent across time, they may not encode the context as strongly, thus resulting in more forgetting of the context relative to the exemplar. Regardless of whether the current pattern of results is due to forgetting or consolidation, this set of experiments is the first to suggest that one process of word learning is learning what to remember and what to forget.

While both the consolidation and forgetting accounts explain the relatively better performance in context generalization over exemplar generalization after a delay, what about the specific finding that high vocabulary toddlers did not generalize at above chance levels to new,

same-shaped exemplars after a delay? Why did we not find evidence for a shape bias, which is typically a robust form of generalization? As mentioned in the introduction, previous research has found that one mechanism behind the shape bias is the saliency of shape relative to other features (Smith, Landau, & Jones, 1992). One consequence of forgetting or consolidation during the time delay may be that shape becomes just another aspect of the referent; shape is not completely forgotten, but its relative association with the referent may indeed become weaker (via forgetting or consolidation) and thus more in line with the strength of other referent features, such as color. It is unclear why this would have occurred only for the high vocabulary group, and thus additional research is needed to explore this finding. However, by examining the shape bias not only across vocabulary size, but also across time, this study reveals that the shape bias may not be as robust if time elapses between training and generalization.

Limitations of the Current Design

While the results extend previous theories on generalization and word learning in an interesting new direction, there are a couple limitations of the current set of experiments that should be considered. Firstly, it is possible that the different pattern of results for exemplar versus context generalization is due to a difference in saliency of the exemplar and context changes in the generalization trials. The context of the referent images takes up a larger area, and thus the change in context may be more salient than the change in exemplar. This could result in better performance in the context condition simply because the change causes the toddlers to focus more on the task at hand. These potential saliency effects may also explain why the low vocabulary group in Experiment 1's exemplar condition did not show above-chance accuracy in the retention trials. They may have not been able to focus on the task due to the potentially more repetitive nature of this condition compared to the context generalization condition. Indeed,

although the replication sample showed above-chance performance, the low-vocabulary group still performed worse than the high vocabulary group in retention across a short delay. Future studies will use an eye-tracker to examine how toddlers allocate attention throughout the task in order to shed light on the attentional factors in both generalization conditions.

Secondly, generalization type (Exemplar vs. Context) and delay were both manipulated between subjects. Since different toddlers participated in each condition, the current studies do not speak to whether individual toddlers' representations are more flexible across context than exemplar features, or whether representations *change* over time. Answering these questions would require within-subjects and longitudinal designs (respectively). Future studies will test exemplar and context generalization within subjects, but the conclusions from these follow-ups must take into account how seeing one type of generalization trial first may result in additional learning; if toddlers see a change in exemplar color initially, they may infer that the consistent aspect of the referent is the context, which may lead to poor generalization across context due to this experience rather than biases that they are brining to the task. Likewise, a longitudinal study examining generalization performance across different delays within the same group of subjects would be more difficult to interpret. Because participants would necessarily be asked to retrieve the novel words after a short delay first, a longitudinal design would confound delay length and past retrieval. When tested after a long delay, participants would have already seen the novel referent with different exemplar or context patterns (in the first generalization test), and thus it would be unclear if changes in performance were due to the delay length or this past generalization experience. However, longitudinal design will be run, bearing this potential confound in mind, to examine changes in individual toddlers' representations across time.

Applications and Future Directions

Despite these limitations, the current study takes the first step in examining how the flexibility toddlers' novel word representations is affected by time delays and productive vocabulary size. The findings add to a growing literature that suggests that in order to fully understand word learning processes, we must not only examine what toddlers encode about novel words, but what they remember about novel words over time (Wojcik, 2013). Specifically, the current experiments demonstrate that one effect of time on newly learned words is that they may become more flexible across contexts, but less flexible across exemplar features. However, the effect of vocabulary suggests that this change in flexibility may be a learned strategy.

The results are particularly striking because past work on context generalization has almost exclusively focused on how variability during learning can lead to better abstraction (for a few relevant examples, see Gogate & Hollich, 2010; Goldenberg & Sandhofer, 2013; Smith & Vela, 2001). A robust finding, across domains, is that if you vary the context (or any feature that should be abstracted over) across learning trials, children and adults will more easily generalize what they have learned to novel contexts at test. The current study suggests an alternative route to better generalization. By learning more about a particular domain, such as words, toddlers may learn what features to remember over time. They may not need to see variability because they already know from past experiences which aspects of the stimuli are important, and which are not.

This new route to generalization has many applications. For example, in math education, one main goal is to teach children to transfer conceptual knowledge from one problem to another, across contexts. Researchers have found that variability in the contextual features of a math problem leads to better transfer (e.g., Paas & Van Merriënboer, 1994). However, the current results suggest that instead of training generalization for individual problems via

variability during learning, it may be more efficient to train students on a broad range of math concepts, so that they learn which features of a new concept should be retained, and which features can be generalized over. Over time, children may be able to learn what to remember, and what can be generalized over, when they are exposed to only one example of a new math concept.[7]

To further understand the characteristics and development of this learning-to-remember strategy, there are several future directions that will be taken. Firstly, I will examine the relationship between productive vocabulary and the flexibility of novel word representations more thoroughly. Does productive vocabulary size indeed result in a better understanding of what characteristics of novel words should be remembered, as suggested? Or is vocabulary size simply an index of a domain-general ability such as attention? Additionally, further research must tease out the causal story. It could be that a better focus on the referent is what caused to some children having better vocabularies, not the other way around. As with the work on the shape bias (Colunga & Smith, 2005; Jones & Smith, 2002; Samuelson & Smith, 1999), longitudinal training studies and computational modeling will help answer these questions.

Secondly, the current study only investigated generalization across exemplar and context color. There are many other visual and auditory features of a word learning moment, and these other features may be encoded and remembered differently. For example, just as exemplar color was inflexible for high vocabulary toddlers, the function of an object may be an inflexible property of novel nouns as well. On the other hand, just as context color was easily generalized over for high vocabulary toddlers, other features of a word learning moment, such as who is holding the novel object or where it is placed, may also become more generalizable. Future

[7] Thank you to Martha Alibali for pointing out this promising application.

studies will use the same methodology as the current experiments to investigate which aspects of a word learning moment are flexible and which are not, both across time and vocabulary development.

Similar methods will also be used investigate the flexibility of novel word labels. As mentioned previously, early lexical representations are specific to individual phonemes: "vaby" is not an acceptable variant of "baby" (Swingley & Aslin, 2000). However, there are other aspects of novel word labels that should be more flexible. Speaker identity, for example, may be encoded by infants and toddlers, but may become easy to generalize over as more labels are learned. Prosody is another acoustic characteristic that must be generalized over. These aspects of novel word labels should be investigated across time delays and development as well.

Lastly, other lexical categories, such as adjectives and verbs, will be examined. Much of the research on generalization, and novel word learning overall, has focused on concrete nouns. It is possible that when toddlers learn verbs or proper nouns, they encode features of the learning moment—such as context—differently due to the constraints of that lexical category. By studying other types of words beyond nouns, we will better understand the factors that affect what children encode and generalize when they hear a new word.

Conclusion

After learning the word "cup", will a toddler extend that word to another cup? To another context? How is this flexibility, a crucial aspect of semantic representations, affected by time and by the number of words that the toddler knows? While previous research has found that young children have difficulty generalizing words to new contexts, the current study demonstrates that toddlers are able to show flexibility across contexts in a more implicit task. Additionally, by testing both one-minute and one-week delays, the current study challenges the assumption that

toddlers' performance in word comprehension tasks immediately after learning reveals the quality of their lexical representations. What toddlers remember about a word over time appears to be affected by how many words they know; only toddlers with higher vocabularies have novel word representations that are specific to the exemplar but flexible across contexts after a delay. As toddlers learn more words, they begin to efficiently focus on remembering the referent, not the background context. This study provides the first piece of evidence that toddlers do not just learn to learn; they also learn to remember.

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.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063.

Baldwin, D. A., & Markman, E. M. (1989). Establishing word-object relations: A first step. *Child Development*, *60*, 381–398.

Ballem, K. D., & Plunkett, K. (2005). Phonological specificity in children at 1;2. *Journal of Child Language*, *32*(1), 159–173.

Barnat, S. B., Klein, P. J., & Meltzoff, A. N. (1996). Deferred imitation across changes in context and object: Memory and generalization in 14-month-old infants. *Infant Behavior and Development*, *19*(2), 241–251.

Bauer, P. J. (2004). Getting explicit memory off the ground: Steps toward construction of a neuro-developmental account of changes in the first two years of life. *Developmental Review*, *24*(4), 347–373.

Bauer, P. J. (2005). Developments in declarative memory decreasing susceptibility to storage failure over the second year of life. *Psychological Science*, *16*(1), 41–47.

Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30months. *Cognition*, *126*(1), 39–53.

Booth, A. E. (2009). Causal supports for early word learning. *Child Development*, *80*(4), 1243–1250.

Booth, A. E., & Waxman, S. R. (2002). Word learning is "smart": Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, *84*, B11–B22.

Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 556.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133.

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. *Developmental Science*, *12*(5), 815–823.

Carey, S., & Bartlett, E. (1978). Aquiring a single new word. In *Papers and Reports on Child Language Development* (Vol. 15, pp. 17–29).

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Cimpian, A., & Erickson, L. C. (2012). Remembering kinds: New evidence that categories are privileged in children's thinking. *Cognitive Psychology*, *64*(3), 161–185.

Cimpian, A., & Markman, E. M. (2005). The Absence of a Shape Bias in Children's Word Learning. *Developmental Psychology*, *41*(6), 1003–1019.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*(2), 1–36.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Leraning and Verbal Behavior*, *11*, 671–684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268.

Davis, M. H., Di Betta, A. M., Macdonald, M. J. E., & Gaskell, M. G. (2009). Learning and Consolidation of Novel Spoken Words. *Journal of Cognitive Neuroscience*, *21*(4), 803–820.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics*, *21*(01), 95–116.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening Using eye movements to monitor spoken language. In I. A. Sekerina, E. M. Fernandez, & H. Clasen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97–135). Amsterdam: John Benjamins.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331.

Gogate, L. J., & Hollich, G. (2010). Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological Review*, *117*(2), 496–516.

Goldenberg, E. R., & Sandhofer, C. M. (2013). Same, varied, or both? Contextual support aids young children in generalizing category labels. *Journal of Experimental Child Psychology*, *115*(1), 150–162.

Gomez, R. L., Newman-Smith, K. C., Breslin, J. H., & Bootzin, R. R. (2011). Learning, memory, and sleep in children. *Sleep Medicine Clinics*, *6*(1), 45–57.

Goodman, J. C., McDonough, L., & Brown, N. B. (1998). The role of semantic context and memory in the acquisition of novel nouns. *Child Development*, *69*(5), 1330–1344.

Haaf, R. A., Lundy, B. L., & Coldren, J. T. (1996). Attention, recognition, and the effects of stimulus context in 6-month-old infants. *Infant Behavior and Development*, *19*(1), 93–106.

Hartshorn, K., Rovee-Collier, C., Gerhardstein, P., Bhatt, R. S., Klein, P. J., Aaron, F., Wurtzel, N. (1998). Developmental changes in the specificity of memory over the first year of life. *Developmental Psychobiology*, *33*(1), 61–78.

Hollich, G., Golinkoff, R. M., & Hirsh-Pasek, K. (2007). Young children associate novel words with complex objects rather than salient parts. *Developmental Psychology*, *43*(5), 1051–1061.

Horst, J. S. & Hout, M. C. (2014). The Novel Object and Unusual Name (NOUN) Database: a collection of novel images for use in experimental research. Unpublished manuscript.

Horst, J. S., & Samuelson, L. K. (2008). Fast Mapping but Poor Retention by 24-Month-Old Infants. *Infancy*, *13*(2), 128–157.

Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language Experience Shapes the Development of the Mutual Exclusivity Bias. *Infancy*, *15*(2), 125–150.

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

Jaswal, V. K., & Markman, E. M. (2003). The relative strengths of indirect and direct word learning. *Developmental Psychology*, *39*(4), 745–760.

Jones, E. J. H., Pascalis, O., Eacott, M. J., & Herbert, J. S. (2011). Visual recognition memory across contexts: Memory across contexts. *Developmental Science*, *14*(1), 136–147.

Jones, S. S., & Smith, L. B. (2002). How children know the relevant properties for generalizing object names. *Developmental Science*, *5*(2), 219–232.

Lany, J., & Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development. *Developmental Science*, *14*(5), 1207–1219.

Learmonth, A. E., Lamberth, R., & Rovee-Collier, C. (2004). Generalization of deferred imitation during the first year of life. *Journal of Experimental Child Psychology*, *88*(4), 297–318.

Mandler, J. M., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition*, *59*(3), 307–335.

Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language*, *57*(2), 252–272.

Mani, N., & Plunkett, K. (2008). Fourteen-month-olds pay attention to vowels in novel words. *Developmental Science*, *11*(1), 53–59.

Maren, S., Aharonov, G., & Fanselow, M. S. (1997). Neurotoxic lesions of the dorsal hippocampus and Pavlovian fear conditioning in rats. *Behavioural Brain Research*, *88*(2), 261–274.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*, 813–815.

Mather, E., & Plunkett, K. (2009). Learning Words Over Time: The Role of Stimulus Repetition in Mutual Exclusivity. *Infancy*, *14*(1), 60–76.

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190–1210.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McGaugh, J. L. (2000). Memory--a Century of Consolidation. *Science*, *287*(5451), 248–251.

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name–nameless category (N3C) principle. *Child Development*, *65*(6), 1646–1662.

Mills, D. L., Plunkett, K., Prat, C., & Schafer, G. (2005). Watching the infant brain learn words: effects of vocabulary size and experience. *Cognitive Development*, *20*(1), 19–31.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R.* Chapman and Hall/CRC Press.

Munakata, Y. (2001). Graded representations in behavioral dissociations. *Trends in Cognitive Sciences*, *5*(7), 309–315.

Munro, N., Baker, E., McGregor, K., Docking, K., & Arculi, J. (2012). Why Word Learning is not fast. *Frontiers in Psychology*, *3*.

Paas, F. G., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122.

Perry, L. K., & Samuelson, L. K. (2011). The Shape of the Vocabulary Predicts the Shape of the Bias. *Frontiers in Psychology*, *2* (345).

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239.

Samuelson, L. K., Schutte, A. R., & Horst, J. S. (2009). The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization. *Cognition*, *110*(3), 322–345.

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*(1), 1–33.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, *34*(7), 1244–1286.

Smith, L. B., Colunga, E., & Yoshida, H. (2010). Knowledge as process: Contextually cued attention and early word learning. *Cognitive Science*, *34*(7), 1287–1314.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Smith, L. B., Landau, B., & Jones, S. S. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology*, *28*(2), 273–286.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203–220.

Spiegel, C., & Halberda, J. (2011). Rapid fast-mapping abilities in 2-year-olds. *Journal of Experimental Child Psychology*, *109*(1), 132–140.

Stickgold, R. (2013). Early to bed: how sleep benefits children's memory. *Trends in Cognitive Sciences*, *17*(6), 261–262.

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, *43*(2), 454–464.

Swingley, D., & Aslin, R. N. (2000). Spken word recognition and lexical representation in very young children. *Cognition*, *76*, 147–166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken lanuguage comprehension. *Science*, *268*(5217), 1632–1634.

Thiessen, E. D., & Yee, M. N. (2010). Dogs, bogs, labs, and lads: What phonemic generalizations indicate about the nature of children's early word-form representations. *Child Development*, *81*(4), 1287–1303.

Vlach, H. A. (2014). The spacing effect in children's generalization of knowledge: Allowing children time to forget promotes their ability to learn. *Child Development Perspectives, 8*(3), 163-168.

Vlach, H. A., Ankowski, A. A., & Sandhofer, C. M. (2012). At the same time or apart in time? The role of presentation timing and retrieval dynamics in generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 246–254.

Vlach, H. A., & Kalish, C. W. (2014). Temporal dynamics of categorization: forgetting as the basis of abstraction and generalization. *Frontiers in Psychology*, *5*.

Vlach, H. A., & Sandhofer, C. M. (2011). Developmental differences in children's context-dependent word learning. *Journal of Experimental Child Psychology*, *108*(2), 394–401.

Vlach, H. A., & Sandhofer, C. M. (2012). Fast mapping across time: Memory processes support children's retention of learned words. *Frontiers in Psychology*, *3*.

Walker, M. P., & Stickgold, R. (2010). Overnight alchemy: sleep-dependent memory evolution. *Nature Reviews Neuroscience*, *11*(3), 218–218.

Waxman, S. R., & Booth, A. E. (2000). Principles that are invoked in the acquisition of words, but not facts. *Cognition*, *77*(2), B33–B43.

Werchan, D. M., & Gómez, R. L. (2014). Wakefulness (not sleep) promotes generalization of word learning in 2.5-year-old children. *Child Development*, *85*(2), 429-436.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, *3*(1), 1–30.

Wiltgen, B. J., & Silva, A. J. (2007). Memory for context becomes less specific with time. *Learning & Memory*, *14*(4), 313–317.

Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia*, *48*(8), 2339–2356.

Winocur, G., Moscovitch, M., & Sekeres, M. (2007). Memory consolidation or transformation: Context manipulation and hippocampal representations of memory. *Nature Neuroscience*, *10*(5), 555–557.

Wixted, J. T. (2004). The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology*, *55*(1), 235–269.

Wojcik, E. H. (2013). Remembering New Words: Integrating Early Memory Development into Word Learning. *Frontiers in Psychology*, *4*.

Wojcik, E.H., Lew-Williams, C., & Saffran, J.R. (2014). Putting words in their place: 18-month-olds' lexical representations are tied to context. Manuscript under revision.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*(4), 553.

Appendix

The following figures show, for all trial types across both experiments and vocabulary levels, the percent of participants who had a mean accuracy above 50% (thus demonstrating above chance performance). Note that the pattern of results with this alternate analysis mirrors the pattern reported in the manuscript.

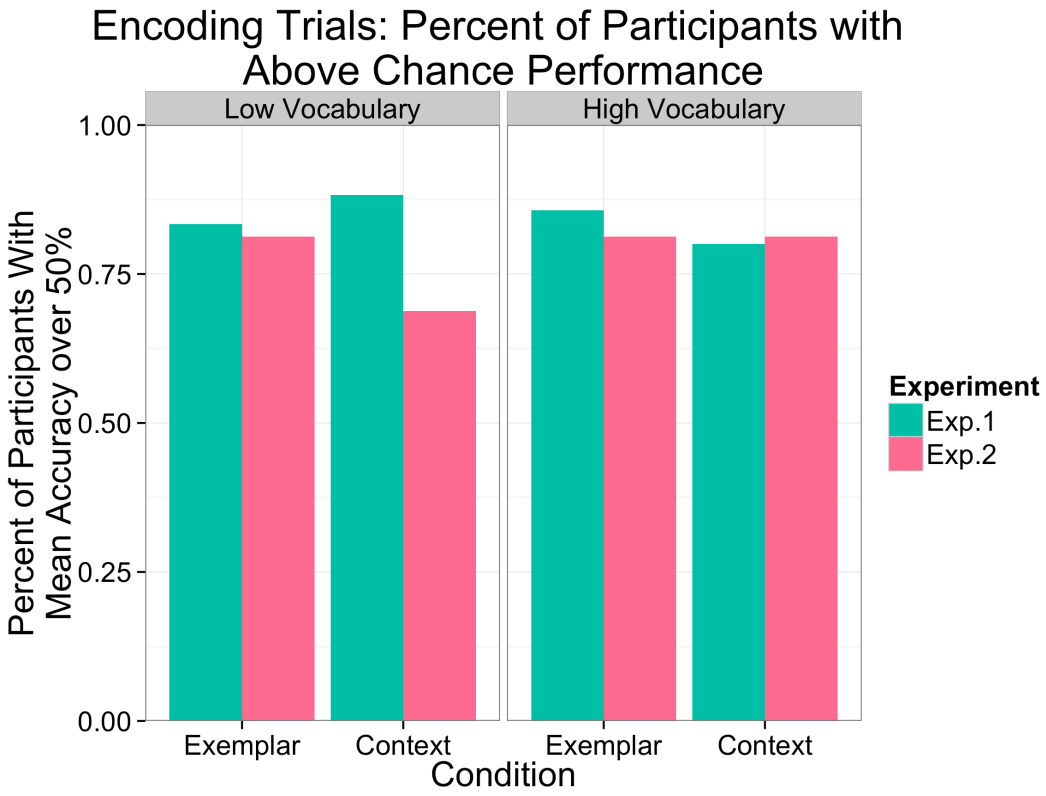

Figure A1. The proportion of participants with above chance performance on the encoding trials for each condition, group, and experiment.

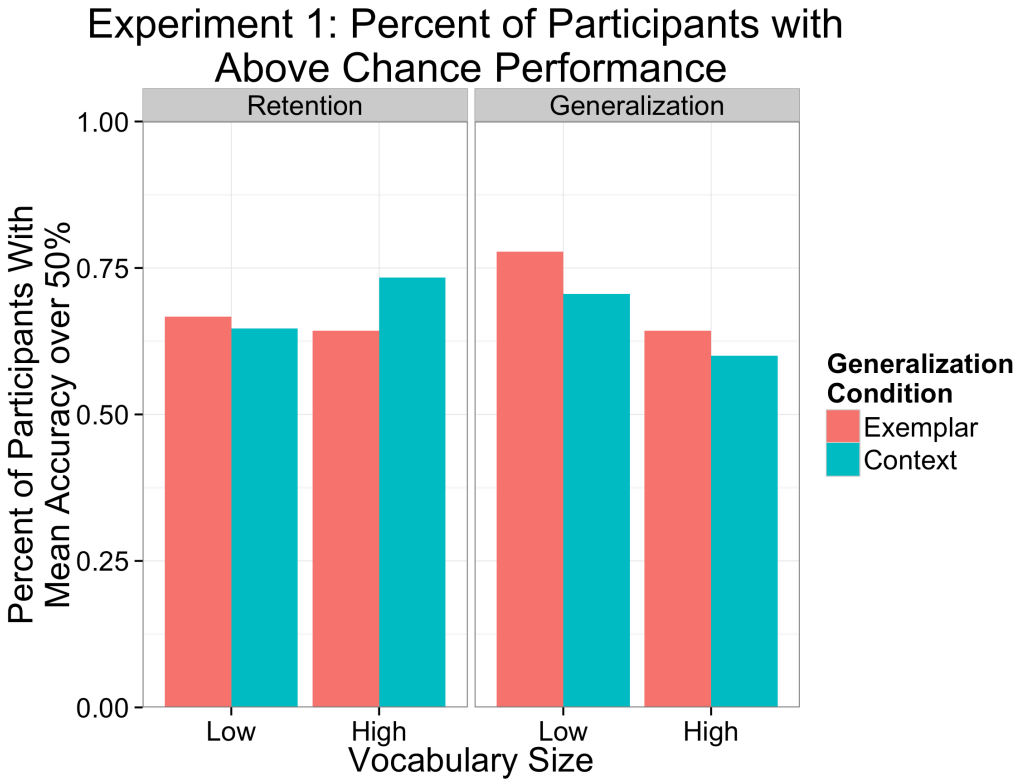

Figure A2. The proportion of participants with above chance performance on the retention and generalization trials for each condition and group in Experiment 1. This figure is analogous to Figure 7.

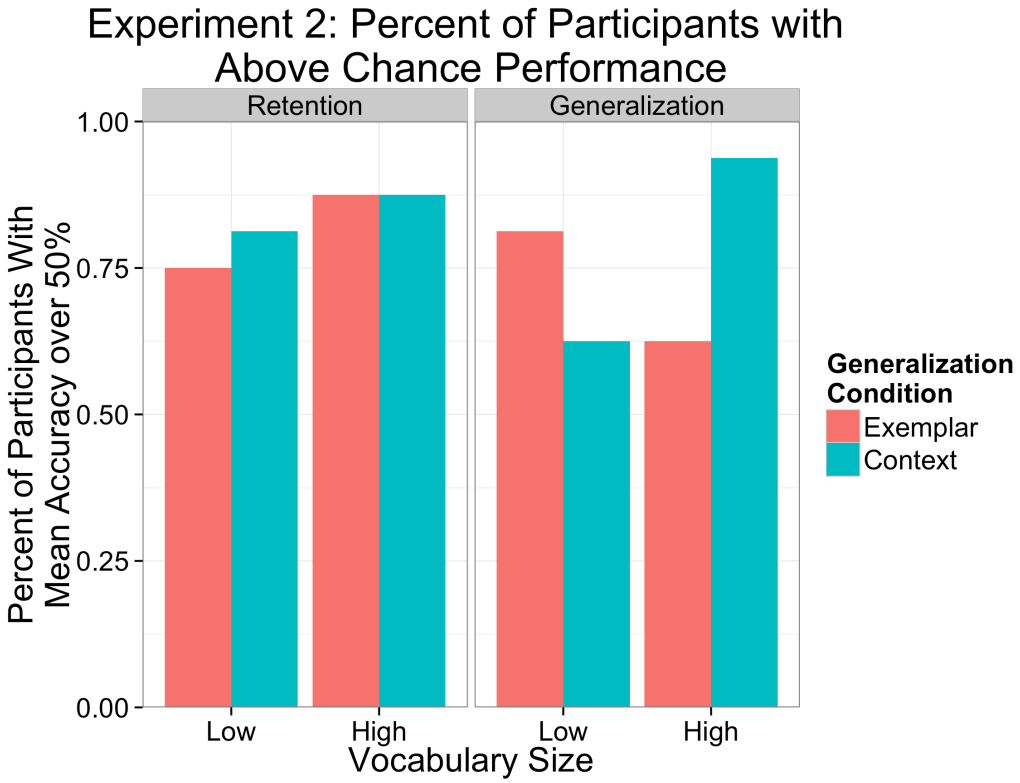

Figure A3. The proportion of participants with above chance performance on the retention and generalization trials for each condition and group in Experiment 2. This figure is analogous to Figure 8.