2017

# 2.5-Year-Olds' Retention and Generalization of Novel Words across Short and Long Delays

Erica H. Wojcik
*Skidmore College*

## Recommended Citation

Erica H. Wojcik (2017): 2.5-Year-Olds' Retention and Generalization of Novel Words across Short and Long Delays, Language Learning and Development

# 2.5-year-olds' retention and generalization of novel words across short and long delays

Erica H. Wojcik

Skidmore College

Department of Psychology
815 North Broadway Street
Saratoga Springs, NY 12866

Phone: 518-580-8306
Email: wojcik.erica@gmail.com

Two experiments investigated two-year-olds' retention and generalization of novel words across short and long time delays. Specifically, retention of newly learned words and generalization to novel exemplars or novel contexts were tested one minute or one week after learning. Experiment 1 revealed successful retention as well as successful generalization to both new exemplars and new contexts after a one-minute delay, with no statistical differences between retention and generalization performance for either generalization type. Toddlers tested after a week delay (Experiment 2) showed successful retention and generalization as well, but while context generalization was statistically equivalent to retention accuracy, exemplar generalization was significantly lower than retention accuracy. The overall success in both retention and generalization suggests that toddlers' newly learned words are robust and flexible. However, the lower exemplar generalization performance compared to retention after a weeklong delay suggests that novel words may become less flexible across exemplar characteristics over time.

(150)

# 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, she may encode the association between that label and that blue cup. Learning this one association is an important step in word learning, but the child must move beyond this specific learning moment. She must be able to remember that word days and weeks later, as well as correctly generalize “cup” to other cups in other contexts. The ability to flexibly understand a word in novel situations across time is a crucial aspect of word learning (see Wojcik, 2013; Colunga & Smith, 2005).

Recent work has explored toddlers’ retrieval of novel words across time, presenting a mixed picture of novel word retention. Two-year-olds have difficulty retaining novel words for even five minutes under some learning conditions (e.g., Bion, Borovsky, & Fernald, 2013), but can retain words for up to 24 hours if the novel object is explicitly labeled (e.g., Goodman, McDonough, Brown, 1998; Horst & Samuelson, 2008). Three-year-olds can retain a novel word for a week or more if the novel label-referent mapping is made explicit and salient (e.g., the object is directly labeled six times and the toddler is prompted to produce the label), but the learned mapping follows a curvilinear pattern of decay over time (Vlach & Sandhofer, 2012). Thus, while it is clear that young children can retain words for long periods of time, two-year-olds’ novel words appear to be fragile and dependent on multiple levels of encoding support during learning.

Interestingly, given the recent investigations into the fragility of two-year-olds’ newly learned words, we know surprisingly little about how the quality of novel word representations change across time. Novel words are more than just a mapping between a label and a referent. Semantic representations must be specific to some dimensions but flexible across others. For concrete count nouns, representations must be specific to shape (cups are cup-shaped), but flexible across many other exemplar and context conditions; a cup is still a cup regardless of its color, or whether it is on the kitchen table or living room floor. Previous research has shown that at 2 to 3 years of age, children are able to generalize novel count nouns to new exemplars that differ in color and material as long as the shape of the object remains the same (e.g., Samuelson & Smith, 1999), suggesting that by this age, representations of count nouns are flexible across these exemplar features. More recently, researchers have begun to examine toddlers' ability to generalize novel words to new visual contexts, finding that until 4 to 5 years of age, children's retrieval of newly learned words is significantly impaired if the context is novel at test (Vlach & Sandhofer, 2011), with some experiments showing that two-year-olds fail to retrieve newly learned words in a novel context (Goldenberg & Sandhofer, 2013). Taken together, the literature suggests that two-year-olds' word representations may be more specific to background context than exemplar color.

At two years of age, though, typically developing children understand thousands of words, and it is thus surprising that they have difficulty with visual context generalization. In order to better understand two-year-olds' novel noun retention and flexibility, the current experiments investigate the effect of time on novel word retention and generalization. Research on word learning and memory representations has shown that novel word representations are fragile and can be quickly forgotten. However, memory research also suggests that the passage of time leads to more flexibility, and thus has the potential to shed light on the retention and generalization of novel words.

Studies with both humans and other animals have demonstrated that as time passes and memories are translated from the hippocampus to cortical areas, they are more easily generalized to novel contexts (Maren, Aharonov, & Fanselow, 1997; McGaugh, 2000; Winocur, Moscovitch, & Bontempi, 2010; Wiltgen & Silva, 2007; see McClelland et al., 1995 for a synthesis and model of this perspective for adult humans). While the current study examines visual background context, there are many types of context that have been examined by language and memory researchers (see Faber & León-Araúz, 2016 for a discussion of the many ways context has been operationalized). For example, the context surrounding the word "cup" could be the linguistic frame around the word, the ambient noise, the visual background etc. The common link between different types of context is that they are all aspects of the input that vary widely across situations and are thus conventionally considered extra-representational. For the current experiments, research on many types of context have been considered because of this commonality; the type of context is stated when necessary for clarity.

Memory studies using a contingency-learning paradigm have found that six-months-olds' memories are only bound to visual context if tested less than a week after learning (Borovsky & Rovee-Collier, 1990). From this literature, one prediction is that after longer delays, younger children will show more robust context generalization of novel words. 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. We know that mature lexical representations are context-specific in some ways; adults are often faster to activate words if they are heard or read in an appropriate (as opposed to novel) linguistic context (Adelman, Brown, & Quesada, 2006; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). It is possible that visual context, like linguistic context, remains an integral aspect of lexical-semantic representations over time.

There is less literature on how time and consolidation affects other types of generalization, such as across exemplar features. It is possible that when toddlers learn novel count nouns, they attend to (and thus encode) shape such that they will use this property to generalize regardless of the delay between learning and test. Recent work has found that for 2-year-olds who have developed a shape bias, memory for the shape of an object is stronger than for its other properties across various delays up to five minutes (Vlach, 2014). Relatedly, shape appears to be a salient feature across a five-minute delay for toddlers with more shape-based nouns in their vocabulary (Perry, Axelsson & Horst, 2015). However, it is possible after a longer delay of multiple days, as a referent representation decays and is consolidated, the saliency of various properties of the referent (such as color, texture, and shape) become more equivalent, leading to less robust shape-based exemplar generalization. This possibility is supported by the drop in novel word retrieval accuracy after a weeklong delay (Vlach & Sandhofer, 2012), which indicates that novel word representations continue to decay, and potentially change, across longer delays.

Despite these interesting predictions from the memory literature, suggesting that time delays may lead to more flexible, generalizable novel words representations (via consolidation), there is surprisingly little work on how toddlers' generalizations change over time. Understanding how toddlers generalize words after a time delay will not only reveal how consolidation affects generalization, but more broadly, it will reveal how the quality of toddlers' semantic representations changes over time. While decades of research have examined retrieval and generalization of newly learned words immediately or a few minutes after learning, very few studies have examined the quality of newly learned words after longer time delays (see Wojcik, 2013 for a review). Infants and toddlers must retrieve words beyond the initial learning moment, and so understanding how word representations are remembered will greatly increase our knowledge of how children build their lexicon. Thus, the aim of this study is to investigate the effect of time on retention as well as on exemplar and context generalization to better understand toddlers' novel word flexibility.

**Comparing Exemplar and Context Generalization**

A secondary aim of the current study is to directly compare exemplar and context generalization after a single-exemplar training phase. One barrier to understanding the factors that lead to the mature generalization of novel words across different features is that dissimilar training and testing protocols are used across generalization studies. In shape bias studies, toddlers are first shown one exemplar for a novel word, and then presented with new objects that differ in either shape, material, or color from the training exemplar (e.g., Landau, Smith, & Jones, 1988; Perry & Samuelson, 2011; Samuelson & Smith, 1999). In context generalization studies, children are exposed to multiple exemplars of a novel word, each on the same colored background, but with different object colors. At test, they are presented with an array of visually distinct objects, including the target, on a new background (e.g., Goldenberg & Sandhofer, 2013; Vlach & Sandhofer, 2011; Werchan & Gómez, 2014). Thus, in context generalization studies, toddlers must abstract over the training exemplars, and then retrieve that representation to select the appropriate object at test.

This difference makes it difficult to compare exemplar and context generalization studies. Specifically, the exemplar variability in the training phase of context generalization experiments introduces the process of abstraction over the stimuli (potentially focusing participants' attention on the invariant context behind the referent; Gogate & Hollich, 2010), which does not exist in exemplar generalization studies. Thus, while the current literature suggests that toddlers are able to generalize novel words to new exemplars two years before they are able to generalize novel words to new contexts (Samuelson & Smith, 1999; Vlach & Sandhofer, 2011)—and thus that early representations are more specific to context than exemplar features—this hypothesis has not been adequately tested.

### Overview of the Current Study

To summarize, the current study aims to examine how time influences novel word retrieval and generalization. The main hypothesis is that novel word representations will be more flexible across context characteristics after a long delay. A secondary hypothesis is that when the same training and testing protocol is used, two-year-olds will be successful at both exemplar generalization and context generalization, but with an interaction with time, reflecting the hypothesis above.

To test these hypotheses, 2.5-year-olds were trained on four novel words, each of which labeled a novel object presented against a distinct visual background (*Exposure* phase). They were then tested on how well they encoded the words (*Encoding Test* phase). This phase ensured that participants learned the novel words equally well across conditions. 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). The short delay of one minute was chosen based on previous word learning studies that have examined retention across shorter delays (e.g., Bion, Borovsky, & Fernald, 2013).

Within each experiment, half of the participants saw test trials in which the colors of the exemplars were different (Exemplar Condition), and half of the participants saw test trials in which the colors of the visual contexts were different (Context Condition). These *generalization* *trials* were intermixed with *retention trials*, in which the visual stimuli were the same as in the Exposure and Encoding Test phases.

## Experiment 1

Experiment 1 used looking behavior to investigate 2.5-year-olds' exemplar and context generalization of novel words shortly after learning. The prediction was that toddlers would be able to both retrieve the novel words, as well successfully generalize them to new exemplars and contexts, but that context generalization would show lower accuracy compared to retention.

### 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, Wisconsin 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). Seven participants were excluded for fussing out before the testing phase (5 in the Exemplar condition 2 in the Context condition) Nineteen additional participants were excluded for not providing sufficient looking data (9 in the Exemplar condition, 10 in the Context condition; see Data Preparation 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. These backgrounds had been used in word-learning tasks with a similar age group. The object and backgrounds were combined with Adobe Photoshop, resulting in four high-resolution pdfs. Each image was paired with one of four novel labels that followed the phonological properties of 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.

---Insert Figure 1 about here---

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. 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 distractor (non-target) object.

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 (also taken from Vlach et al., 2012; see Figure 1). 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 1). 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 in order to assess the flexibility of the representations across exemplar and context changes. 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 + background 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, both against a neutral grey background), 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 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 simultaneously, with one on the bottom left of the screen and one on the bottom right. The objects were yoked into two pairs such that each target was always paired with the same distractor (i.e., in one counterbalanced condition, the “coodle” referent was always tested against the “tulver” referent). 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 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; the target object was positioned each side an equal number of times. After the test trials, the participants watched an unrelated movie for one minute to maintain attention across the delay. The movie was a song clip from Sesame Street that was chosen for its ability to re-engage the participants and because it did not present any objects that may have served as reminders of the novel words.

***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. 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), to measure of expressive vocabulary (100-word checklist).

**Data preparation.** Looking behavior was coded in 33ms 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, only the first block (eight trials) was used in the analysis[1]. Trials were excluded if the participant was looking away from the target and distractor for 500ms or more of the 1500ms critical window. This critical window was pre-defined as 300 ms after the onset of the word label (in order to adjust for the time it takes to plan an eye movement) to 1800ms after the onset (Fernald et al., 2008).

[1] 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. For the Exemplar Generalization condition, there was a 20% decrease in the amount of usable trials from Block 1 to Block 2, and in the Context condition, there was a 15% decrease. For Experiment 2, there were similar decreases from Block 1 to Block 2: 20% for the Exemplar condition and 18% for the context condition. 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.

Participants were only included if at least half of their trials were usable for each testing phase.

Mean accuracy scores were then calculated for each subject by condition. More specifically, for each trial, the proportion of time spent looking to the target image during the critical window was calculated. 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 the analyses on the *all-words dataset*.

To create a more conservative dataset that ensures performance was not inflated or deflated by words that were not learned in the first place, a *learned-words dataset* was created for the Generalization Phase trials, only including words that were successfully encoded (dthank you to an anonymous reviewer for this suggestion). Specifically, the words for which a given participant did not show above-chance accuracy (0.50) during the Encoding Test removed from the Generalization Test Phase dataset. The average number of words that were successfully encoded per participant was not statistically different between Conditions, t(51)=0.51, n.s. (Exemplar mean=3.25/4, Context mean =3.26/4), and the reduction in sample size due to participants not contributing data to both the retention and generalization trials can be seen in Table 1.

To examine the role of vocabulary size, participants were assigned to low and high vocabulary groups based on  productive vocabulary scores (from parental report; MacArthur-Bastes Communicative Development Inventory: Short Form Level II; Fenson et al., 2000) 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. In the exemplar condition, there were 18 low vocabulary and 14 high vocabulary participants. In the context condition, there were 17 low vocabulary and 15 high vocabulary participants.

**Results**

To first investigate how well participants learned the novel words, descriptive statistics and one-tailed t-tests against chance for the Encoding trials were calculated (Table 1). For each generalization condition, descriptive statistics and comparisons to chance suggest successful word encoding. Additionally, a 2 (Condition: Exemplar vs. Context, between-subjects) X 2 (Vocabulary size: low vs. high, between-subjects) ANOVA was run with mean encoding accuracy as the dependent variable, finding no significant main effect of Condition (F([1,60]=0.44) or Vocabulary (F[1,60]=0.076) and no significant interaction (F[1,60]=2.73). Thus, participants in all conditions showed successful and similar encoding of the novel words. Because there was no significant effect of vocabulary size in these and following analyses, vocabulary size is excluded from all following models for simplicity.

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. Descriptive statistics, comparisons to chance (see Table 1), and time-course graphs (Figure 2), reveal successful and similar performance across conditions and trial types. Indeed, Figure 2 shows that participants in both conditions and trials types hover around chance before the word onset, then increase looking to the target with similar slopes after word onset. To statistically compare performance across Generalization Condition (Exemplar vs. Context, between-subjects) and Trial Type (Retention vs. Generalization trials, within-subjects), 2x2 mixed ANOVAs were run. The *all-words dataset* showed no significant main effects or interactions. For Condition: F(1,62) = 0.014; for Trial Type F(1,60) = 0.472); for Condition x

Trial Type: F(62)=0.201. This same pattern was reflected in the *learned-words dataset*: there was no main effect of Condition (F[1,51] = 0.904.) or Trial Type (F[1,51] = 0.090), as well as no significant interaction (F[1,51] = 1.106; see Figure 2).

--Insert Table 1 about here--

--Insert Figure 2 about here--

**Discussion**

In Experiment 1, participants were able to learn four novel words, and they successfully remembered the words across the one-minute delay. The first key finding from Experiment 1 is that in line with the hypothesis, participants successfully generalized the novel words to both new exemplars and contexts (see Table 1 and Figure 2). Indeed, there was no significant difference in accuracy between the retention and generalization trials for either type of generalization, as demonstrated by the non-significant effect of Trial Type or interaction between Trial Type and Condition for both the all-words and learned-words datasets (see Figure 2). 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 compared to the retention 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), 18- to 23-month-old infants are 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 for this age group 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.

These results deviate from recent work on toddlers' context generalization of newly learned words (Goldenberg & Sandhofer, 2013; Vlach & Sandhofer, 2011), which found a decrease in accuracy if background context was changed at test. However, this was the first study to directly compare context generalization to exemplar generalization with the same training and testing protocol. Specifically, in the current experiment, the object exemplar was not variable during training, as was the case in past context generalization protocols. Thus, participants were not given input suggesting that context was one of the invariant associates of the label (see Gogate & Hollich, 2010). This may have led to more flexibility across background context. By providing participants with the same exposure phase in both the exemplar and context conditions, confounding variables were eliminated to reveal similar and successful generalization.

The context generalization success found in the current experiment is in line with research on infant memory development. Young infants have difficulty retrieving a memory when the background 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.

Notably, 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 and domains (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, 2016). What the current results point to, though, is that by two and a half years of age, toddlers are able to ignore visual contextual changes if necessary in order to find the referent of a newly learned word.

The results from Experiment 1 show that toddlers' newly learned word representations are flexible across context and exemplar changes. Experiment 2 aimed to shed light on how novel word representations change over time by testing a new sample of participants with a one-week delay between training and test.

## Experiment 2

In Experiment 2, a weeklong delay was introduced between the encoding and generalization tests. This delay length was chosen based on previous studies that have examined novel word recognition across long time scales (Booth, 2009; Markson & Bloom, 1997), and because the most significant drop in toddlers' retention of novel words is seen after a week (Vlach & Sandhofer, 2012). The original prediction was that context generalization performance would improve after a longer delay. However, because of the successful context generalization performance in Experiment 1, the revised prediction was that context generalization would continue to be successful after a longer delay.

Experiment 2, more broadly, tests how the strength and specificity of novel words change over time. There is a paucity of work on how novel words are remembered over long time periods (see Wojcik, 2013).  Do new words remain robust after a one-week delay? Do they remain flexible, or are more cues needed to retrieve the representation? The results from the Experiment 2 shed light on this question as well.

## Method

A new sample of 2.5-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 final 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 from the same population and with the same eligibility requirements as in Experiment 1. 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. Seven were excluded for not making it to the testing phase during the first visit (one in the Exemplar condition, six in the Context condition; all remaining participants sat through the second visit). Sixteen participants were excluded for not providing sufficient looking data (Exemplar condition: 3 for visit 1, 3 for visit 2; Context Condition: 5 for visit 1; 5 for visit 2; see Experiment 1 Data Preparation for exclusion criteria), and five for failing to return for the second session.

**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 five minutes total). Caregivers then filled out the MCDI (Fenson et al., 2000). Participants were brought back to the lab a week later. A 3- to 11-day range was used—one week plus or minus four days—to minimize attrition. Many studies examining effects of retention or consolidation insert a 12- or 24-hour delay (e.g., Waxman & Booth, 2000; Werchan & Gomez, 2014), and thus the minimum of three days is still appropriate for the current research questions.  The mean and mode of the delay length in both conditions was 7, and 46 of the 64 participants in the final sample had a delay length of a week or longer—21 in the exemplar condition and 25 in the context condition. The second lab session consisted of just the Generalization Test phase (~2.5 minutes).

**Data preparation.** 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%. Data were prepared for the all-words and learned-words datasets with the same protocol as Experiment 1. In the learned-words dataset, the number of encoded items per participant was not significantly different between Conditions, t(47) = 1.30, n.s. (Exemplar mean=3.08/4, Context mean=3.35/4), and the reduction in sample size due to participants not contributing data to both the retention and generalization trials can be seen in Table 2.

**Results**

Descriptive statistics and comparisons to chance suggest similar and successful novel word encoding across generalization conditions (see Table 2). There was no significant difference in accuracy between the groups in the Encoding Phase, t(62) = 0.41.

--Insert Table 2 about here—

--Insert Figure 3 about here--

As in Experiment 1, the critical question was how well participants generalized the novel words, but in the current experiment, the delay was one week instead of one minute. In both datasets, all groups and trial types showed above-chance performance and increased looking to the target after word onset (Table 2, Figure 3). However, unlike in Experiment 1, the mixed ANOVAs showed different patterns of results for the all-words and learned-words datasets. In the all-words dataset, there were no significant main effects or interactions: For Condition, F(1,62)=0.056, n.s.; for Trial Type, F(1,62)= 1.712, n.s.; for Condition x Trial Type, F(1,62) = 0.928, n.s. In the learned-words dataset, the 2x2 mixed ANOVA showed no significant effect of Condition (F[1,47]= 0.96) or Trial Type, (F[1,47]=1.96), but a significant interaction between Trial Type and Condition: F(1,47)=6.085, p<0.02 (see Figure 3). Follow-up paired-tests were run on each Condition (Exemplar and Context) to examine the form of this interaction.

While there was no significant difference between retention and generalization for the Context Condition, t(22)= 0.91, there was a significant difference for the Exemplar Condition, t(25)=2.54, p<0.02, such that Retention performance (M=0.80, SE=0.035) was significantly higher than Generalization performance (M=0.63, SE=0.046). By looking at the means for the Exemplar Condition in the all-words dataset (i.e., including words that were not successfully encoded; Table 2), it can be seen that this effect was only revealed in the learned-words dataset because retention accuracy was considerably deflated when words that were not learned were included: the Exemplar Retention mean for all-words dataset was 0.09 lower than for the learned-words dataset, but the Exemplar Generalization means for the datasets were the same[2].

To compare retention and generalization behavior across delays, a mixed 2 (Condition) x 2 (Trial Type) x 2 (Delay) ANOVA was run on the learned-words dataset from both experiments. There was no main effect of Condition (F[1, 98]=0.002) or Trial Type (F[1,98]=0.55), nor were any of the two-way interactions significant. However, there was a

[2] This finding provides strong support for filtering data by encoding accuracy in studies that examine novel word retrieval across delays. Most experiments that test retention after a delay include all trials, regardless of encoding performance (although researchers may perform analyses to examine the correlation between encoding and retention scores, e.g., Bion, Borovsky, & Fernald, 2013; Goodman, McDonough & Brown, 1998) or do not test encoding after the training phase at all (e.g., Vlach & Sandhofer, 2012). For 7 of the 8 conditions in the current experiments, performance was numerically lower in the all-words dataset, in which all words were included regardless of whether or not they were learned. This suggests that we may be underestimating group retrieval performance when we do not filter our data by whether or not novel words were encoded.

significant effect of Delay (F[1,98]=4.13, p <0.05) and a significant three-way interaction (F[1,98]=6.24, p<0.02). The significant effect of delay was driven by the fact that overall, participants were more accurate during Experiment 2 (M=0.71, SE=0.023) than Experiment 1 (M=0.64, SE=0.023).

To further investigate the three-way interaction, 2 (Delay) x 2 (Trial Type) ANOVAs were run for each generalization condition. For the Context condition, there were no significant main effects (Delay: F[1,48]=1.814; Trial Type: F[1,48]=0.033) and no significant interaction (F[1,48]=0.982). However, for the exemplar condition, while the main effects were not significant, (Delay: F[1,50]=2.57; Trial Type: F[1,50]=1.42), there was a significant crossover interaction of Trial Type x Delay, F(1,50)=6.40, p<0.02. Follow-up t-tests revealed that the interaction was driven by a significant increase in retention trial accuracy from Experiment 1 (M=0.61, SE=0.050) to Experiment 2 (M=0.79, SE=0.035), t(50)=3.064, p<0.004, and a non-significant decrease in generalization from Experiment 1 (M=0.67, SE=0.048) to Experiment 2 (M=0.63, SE=0.046), t(50)=0.68.

**Discussion**

Experiment 2 tested the hypothesis that a time delay would led to better context generalization performance. However, because of the successful context generalization in Experiment 1, the revised prediction was that context generalization would still be successful after a weeklong delay. Indeed, there were no differences between retention and generalization for the Context generalization condition in either the all-words or learned-words dataset, suggesting that as in Experiment 1, participants' representations were flexible across context characteristics.

Performance in the Exemplar condition, however, suggests that time delays may affect exemplar and context generalization differently. The analysis on the learned-words dataset for the exemplar condition revealed that generalization accuracy was significantly lower than retention accuracy. The fact that participants' looking accuracy was affected by changes to exemplar color, but not by changes to context color, suggests that as time passes, novel word representations retain their flexibility to new contexts while becoming more reliant on the encoded exemplar characteristics for retrieval.

While the comparisons across delays found that participants were significantly better at Exemplar retention and only numerically worse at Exemplar generalization in the weeklong compared to the one-minute delay condition, the overall pattern of results supports the interpretation that exemplar generalization is hindered after a long delay. In the learned-words dataset, as reflected in the significant main effect of Delay, all trial types and conditions show numerically higher scores in Experiment 2 than in Experiment 1, except for Exemplar Generalization (see Tables 1 and 2). This overall increase in accuracy may be due to the fact that the Generalization Phase trials came after several minutes of training and encoding test trials in the short delay condition, but at the beginning of the second test session in the long delay condition. Participants in the short delay condition may have been less engaged in the task, and thus the accuracy drop may be due to task demands instead of the strength of the novel words representations[3]. If this is the case, then the non-significant decrease in generalization performance in the Exemplar condition is a notable outlier in the Experiment 2 results. Future

[3] One difference between the two experiments is that participants in Experiment 2 never saw the one-minute sesame street video that served as an attention getter in Experiment 1 between the encoding and generalization test phases. However, because the video was chosen to have no content in common with the novel words, it is unlikely that the difference explains the results.

studies should investigate whether exemplar generalization does indeed become less robust after time delays.

The overall high and above-chance performance on the retention trials suggests that two-year-olds were able to retain the novel words for a week. This 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 indicated the correct referent at test after a one-week delay. 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. 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 notable difference between the current study and others is the number of novel words. While participants in the current experiments were taught four novel words, some experiments examining retention have used six (Booth, 2009; Vlach & Sandhofer, 2012) or eight novel words (Horst & Samuelson, 2008, Experiment 1A). On the other end of the spectrum, other experiments have only taught one or two novel words (Bion, Borovsky, & Fernald, 2013; Horst & Samuelson, 2008, Experiment 1B). It is possible that the successful retention found in Experiment 2 is due to the relatively small number of novel words. Because of the variation in the literature, future studies are needed to directly test the relationship between the number of novel words and retention performance.

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

## General Discussion

The current study investigated the effect of time on novel word retention and generalization. A secondary aim was to directly compare context and exemplar generalization after a single-exemplar learning moment. To address these aims, toddlers were tested on retention and either exemplar or context generalization after seeing one exemplar of four novel words, with a delay of either one minute (Experiment 1) or one week (Experiment 2) between the learning and test phases. Toddlers accurately retrieved novel words and generalized them to both new exemplars and contexts after a one-minute delay. Indeed, there was no significant difference between performance on retention and generalization trials. Toddlers were also able to retain and generalization words to new exemplars and contexts after a one-week delay. However, participants performed significantly worse on generalization than one-week retention in the Exemplar Condition, but not the Context Condition.

These results differ from the hypotheses derived from the literature on novel word retention and generalization. First, retention performance was strong for toddlers across both delay lengths. Recent research has called into question two-year-olds' memory for novel words, even for five minutes beyond the learning moment (see Bion, Borovsky, & Fernald, 2013; Horst & Samuelson, 2008; Vlach & Sandhofer, 2012). However, the experiments that show retention failure required participants to infer the label-referent mapping via mutual exclusivity or pragmatic context. Indeed, when given multiple types of support (such as six label repetitions or production prompts), three-year-olds can retain words for up to a month (Vlach & Sandhofer, 2012). The current study rests between inference and supported learning—participants heard the novel referents labeled four times each, but learning was passive and non-social. Thus, this study demonstrates that two-year-olds can retain words for at least a week from a moderately supported learning environment.

Performance on the generalization trials was unpredicted as well. Research on memory and consolidation demonstrates that representations become more context-independent with time (McGaugh, 2000), and thus the prediction was that toddlers would show stronger context generalization after a weeklong delay. This was not confirmed, perhaps due to the strong performance in context generalization after a short delay. However, the fact that exemplar generalization performance was lower than retention performance after a weeklong delay, but not a one-minute delay suggests that context generalization was protected across the longer delay, while exemplar generalization was not. Over time, novel word representations are more flexible across context changes than exemplar characteristic features. One possible explanation for this pattern of results comes from recent research on the role of consolidation in semantic integration. Older children and adults only show phonological priming between newly learned words and familiar words after consolidation (Davis, Di Betta, Macdonald, & Gaskell, 2009). The current study may be a semantic analog of this lexical effect: consolidation may have led to more integration with those toddlers' semantic networks, leading to a retrieval pattern that better reflects the characteristics of their lexical-semantic network—specifically, a focus on the referent over the context. Future work will manipulate consolidation directly (via sleep, for example; Walker & Stickgold, 2010) to test this possible mechanism.

In regards to the secondary aim of the study—to compare exemplar and context generalization with the same design—the current findings suggest that contrary to previous research, two-year-olds can successfully generalize newly learned words to novel contexts;

indeed participants' accuracy was not affected by a change to context in either experiment. In past studies, toddlers may have used information during learning to determine what features of the scene were important for the novel word's meaning; since the context was consistent but the exemplar color variable (Vlach & Sandhofer, 2011; Goldenberg & Sandhofer, 2013), context became a robust component of that word's representation. In the current study, toddlers generalized equally well when either the exemplar color or the context color was changed from a single training exemplar, suggesting that by two years of age, children are able to flexibly extend novel words across these two dimensions.

A last finding was that there were no significant effects of vocabulary on retention or generalization accuracy. This finding may be surprising given the role of two-year-olds' vocabulary size in retention and exemplar generalization (e.g., Bion, Borovsky, & Fernald, 2013; Gershkoff-Stowe & Smith, 2004; Mervis & Bertrand, 1994). However, many referent selection tasks do not find effects of vocabulary size (e.g., Horst & Samuelson, 2008; O'Doherty et al., 2011), suggesting that vocabulary size may not affect this type of task. Another possible explanation is that the vocabulary measure used, the MCDI Short Form (Fenson et al., 2000), is not sensitive enough to capture individual differences in vocabulary size, and that a vocabulary size effect would be detected if more comprehensive measure were used, such as the long form MCDI (Fenson et al., 1993).

It is worth noting that planned follow-up analyses on the all-words dataset of Experiment 2 (one-week delay) did reveal that for the high vocabulary group, but not low vocabulary group, context generalization performance was significantly higher than exemplar generalization performance. However, this effect did not stand up to further scrutiny; there was no significant interaction between vocabulary and generalization type when participants were split into three vocabulary groups or when vocabulary was treated as a continuous variable (as suggested by an anonymous review). Additionally, this vocabulary effect was not found when non-encoded words were removed for the learned-words dataset. Because of the lack of support for the vocabulary interaction, further research is needed before conclusions can be drawn about how vocabulary size influences novel word generalization across time.

## Conclusions

The current experiments examined how the retention and flexibility of toddlers' novel word representations is affected by time delays. Specifically, the results demonstrate that two-year-olds' retain and generalize words well not only across a one-minute delay, but also after a weeklong delay. However, novel word representations are more specific to exemplar characteristics than context characteristics after a weeklong delay only. If a toddler is first taught the word "cup" in relation to a blue sippy cup, after a week, she will more easily extend that new word to new contexts than to new cups. 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.

## Acknowledgements

I would like to thank the participating families, as well as the members of the Infant Learning Lab for their help. In particular, I would like to thank Hilary Stein, Erika Steinbauer and Ellen Breen for coding participant data; Rachel Raczynski for helping to create the stimuli; and Erin Long and Courtney Thom for participant recruitment and data collection. I also thank Jenny Saffran, Mark Seidenberg, Tim Rogers, and Haley Vlach for comments on previous versions of this manuscript, and Padmapriya Kandhadai for helpful discussions. This work was funded by a National Science Foundation Graduate Research Fellowship to EHW and by grants from the National Institute of Child Health and Human Development to the Waisman Center (P30HD03352).

## 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.

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.

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

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

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.

Borovsky, D., & Rovee‐Collier, C. (1990). Contextual constraints on memory retrieval at six months. *Child Development*, *61*(5), 1569-1583.

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

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

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.

Faber, P., & León-Araúz, P. (2016). Specialized knowledge representation and the parameterization of context. *Frontiers in psychology*, *7*.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). The MacArthur Communicative Development Inventories: User's guide and technical manual. San Diego: Singular Publishing Group.

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.

Gershkoff‐Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child Development*, *75*(4), 1098-1114.

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.

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

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.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128-157.

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.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299–321.

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*(6619), 813-815.

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*, 1646-1662.

O’Doherty, K., Troseth, G. L., Shimpi, P. M., Goldenberg, E., Akhtar, N., & Saylor, M. M. (2011). Third‐Party Social Interaction and Word Learning From Video. *Child Development*, *82*(3), 902-915.

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

Perry, L. K., Axelsson, E. L., & Horst, J. S. (2015). Learning What to Remember: Vocabulary Knowledge and Children's Memory for Object Names and Features. *Infant and Child Development*.

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.

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

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.

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.

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.

Vlach, H. A. (2014). *The shape bias shapes more than just attention: relationships between categorical bias & object recognition memory.* In Proceedings of the Thirty-sixth Annual Conference of the Cognitive Science Society. Quebec City.

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., & 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). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development*, *83*(4), 1137-1144.

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.

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.

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. (2016). Putting words in their place: 18-month-olds' lexical representations are tied to context. Manuscript under revision.

Table 1

Experiment 1: Mean accuracies across Condition and Trial Type

| Condition | Trial Type | Dataset | *Mean* | *SD* | *df* | *t* | *p* | *Cohen's D* |
|---|---|---|---|---|---|---|---|---|
| Exemplar | Encoding | - | 0.63 | 0.12 | 31 | 5.93 | <0.0001 | 1.08 |
| | Retention | All-words | 0.57 | 0.21 | 31 | 1.79 | 0.041 | 0.33 |
| | | Learned-words | 0.61 | 0.25 | 25 | 2.24 | 0.017 | 0.44 |
| | Generalization | All-words | 0.64 | 0.21 | 31 | 3.60 | <0.0001 | 0.67 |
| | | Learned-words | 0.67 | 0.24 | 25 | 3.62 | 0.00066 | 0.71 |
| Context | Encoding | - | 0.65 | 0.13 | 31 | 6.50 | <0.0001 | 0.63 |
| | Retention | All-words | 0.61 | 0.25 | 31 | 2.41 | 0.012 | 0.44 |
| | | Learned-words | 0.66 | 0.24 | 26 | 3.60 | 0.00066 | 0.67 |
| | Generalization | All-words | 0.59 | 0.20 | 31 | 2.40 | 0.0093 | 0.45 |
| | | Learned-words | 0.63 | 0.23 | 26 | 2.97 | 0.0031 | 0.57 |

*Note.* The all-words dataset includes trials for all words, while the learned-words dataset only includes trials for successfully encoded words (see Experiment 1 Methods for details). Some participants did not contribute retention and generalization trials for the subset of words that they successfully encoded, thus the slight reduction in Ns per cell for the learned-words dataset. Means, standard deviations, degrees of freedom, t-values, and p-values (one-tailed) are for comparisons against chance (0.50).

Table 2

Experiment 2: Mean accuracies across Condition and Trial Type

| Condition | Trial Type | Dataset | *Mean* | *SD* | *df* | *t* | *p* | *Cohen's D* |
|---|---|---|---|---|---|---|---|---|
| Exemplar | Encoding | - | 0.61 | 0.14 | 31 | 4.60 | <0.0001 | 0.79 |
| | Retention | All-words | 0.71 | 0.18 | 31 | 6.49 | <0.0001 | 1.17 |
| | | Learned-words | 0.80 | 0.25 | 25 | 8.40 | <0.0001 | 1.20 |
| | Generalization | All-words | 0.63 | 0.22 | 31 | 3.18 | 0.0017 | 0.59 |
| | | Learned-words | 0.63 | 0.23 | 25 | 2.77 | 0.0052 | 0.56 |
| Context | Encoding | - | 0.63 | 0.16 | 31 | 4.48 | <0.0001 | 0.81 |
| | Retention | All-words | 0.68 | 0.20 | 31 | 5.20 | <0.0001 | 0.90 |
| | | Learned-words | 0.69 | 0.21 | 22 | 4.18 | 0.00020 | 0.90 |
| | Generalization | All-words | 0.67 | 0.24 | 31 | 4.08 | 0.00015 | 0.71 |
| | | Learned-words | 0.74 | 0.26 | 22 | 4.45 | 0.00010 | 0.92 |

*Note.* The all-words dataset includes trials for all words, while the learned-words dataset only includes trials for successfully encoded words (see Experiment 1 Methods for details). Some participants did not contribute retention and generalization trials for the subset of words that they successfully encoded, thus the slight reduction in Ns per cell for the learned-words dataset. Means, standard deviations, degrees of freedom, t-values, and p-values (one-tailed) are for comparisons against chance (0.50).

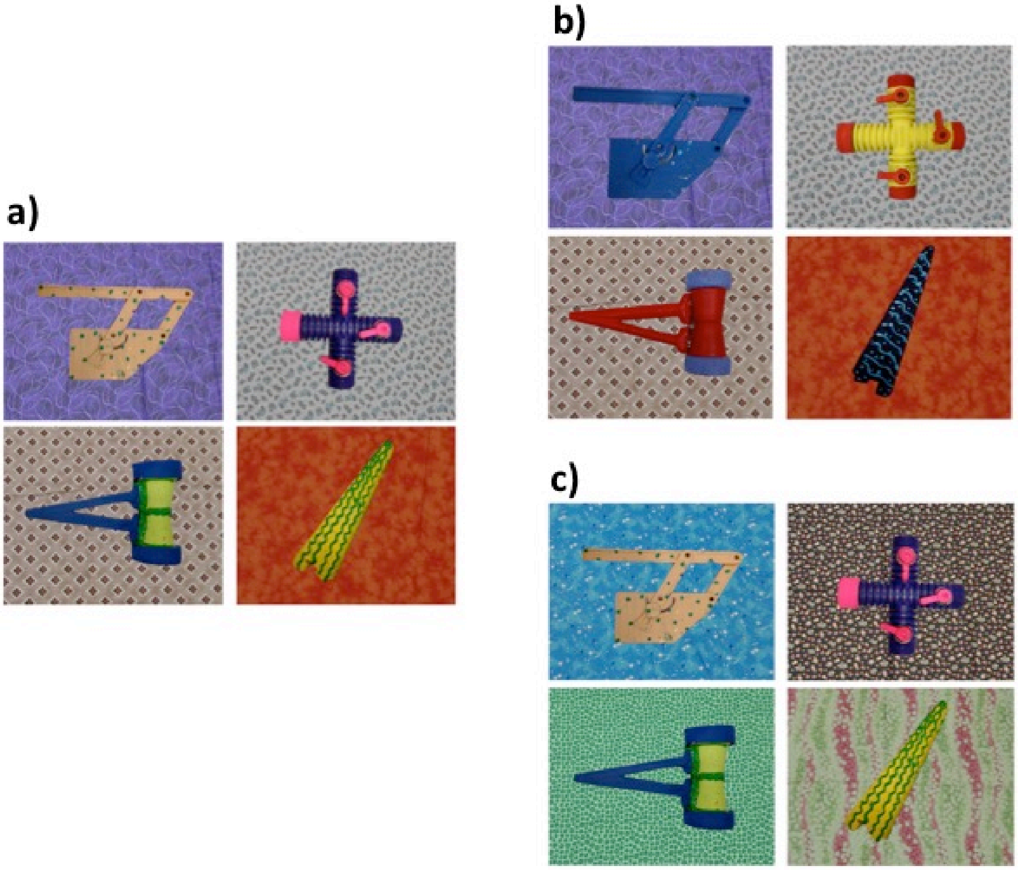

*Figure 1.*The novel object images presented in the a) Encoding phase b) Exemplar Generalizatoin condition and c) Context Generalization condition.

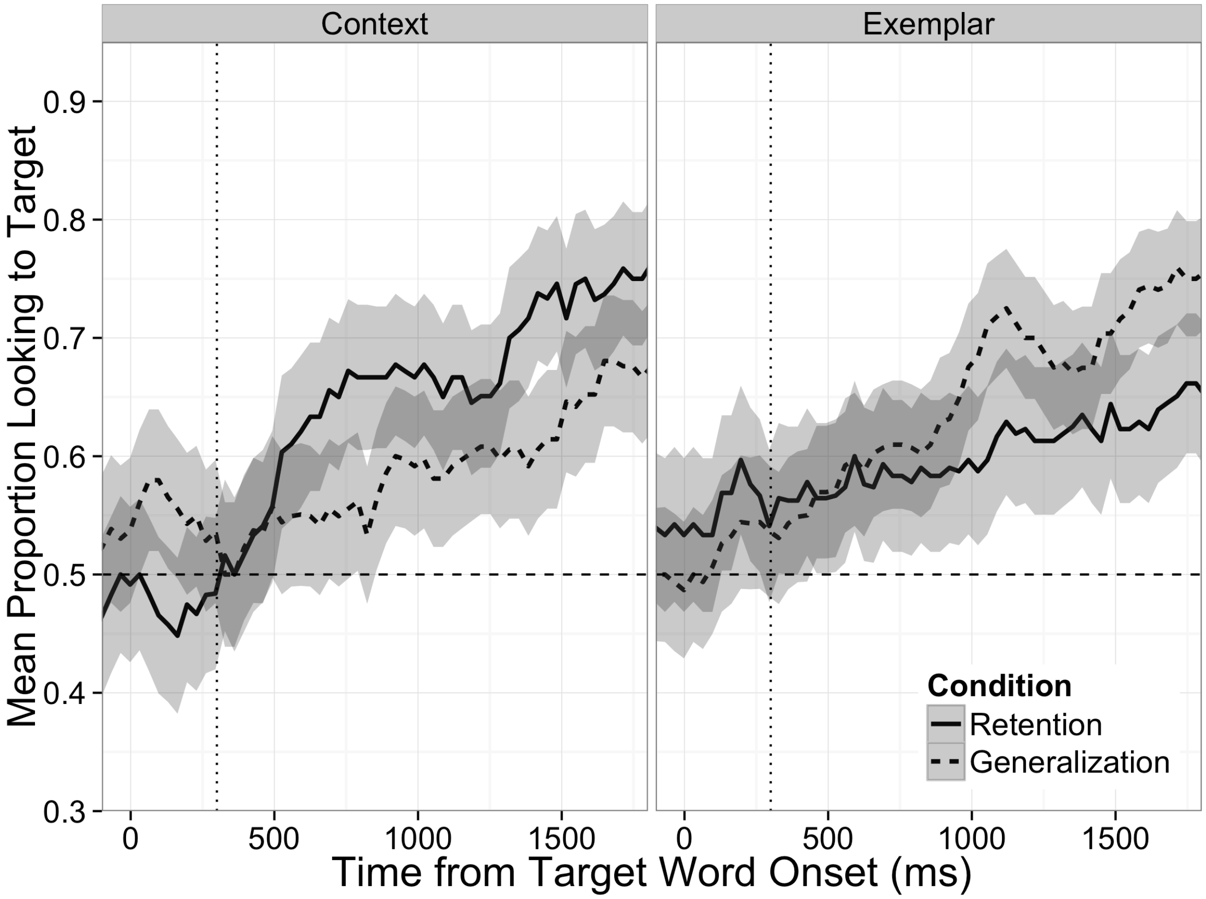

*Figure 2.* Mean proportion of looking to the target image after target word onset in Experiment 1 (learned-words dataset). The dashed horizontal line represented chance performance. The dotted vertical line marks the onset of the critical window (300ms after word onset) used in the analysis. Error bands represent one standard error of the mean.

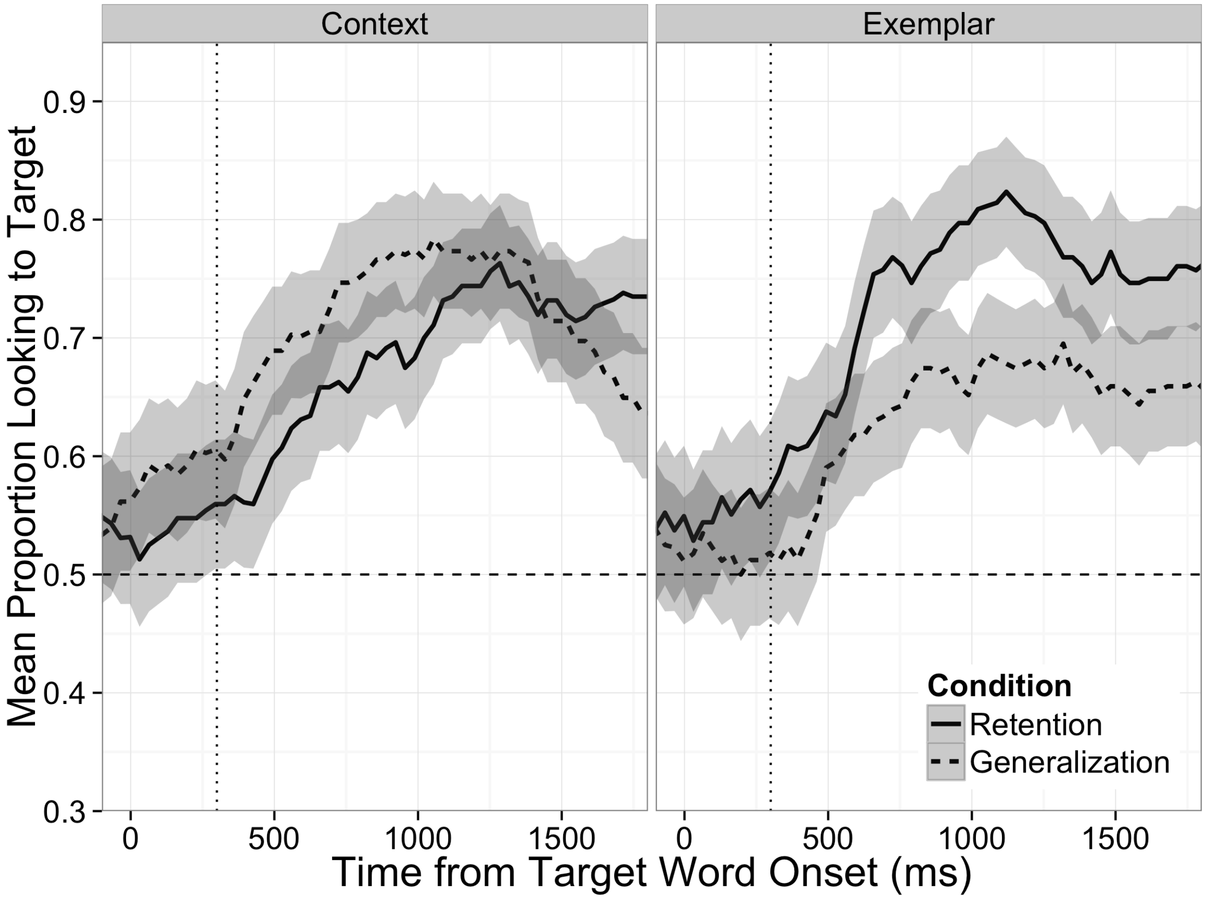

*Figure 3.* Mean proportion of looking to the target image after target word onset in Experiment 2 (learned-words dataset). The dashed horizontal line represented chance performance. The dotted vertical line marks the onset of the critical window (300ms after word onset) used in the analysis. Error bands represent one standard error of the mean.