DOI: 10.1111/cdev.14120

CHILD DEVELOPMENT

REGISTERED REPORT

Becoming word meaning experts: Infants' processing of familiar words in the context of typical and atypical exemplars

Haley Weaver[1,*] | **Martin Zettersten**[2,*] | **Jenny R. Saffran**[1]

[1]Department of Psychology, University of Wisconsin-Madison, Madison, Wisconsin, USA

[2]Department of Psychology, Princeton University, Princeton, New Jersey, USA

Correspondence
Martin Zettersten, Department of Psychology, Princeton University, Princeton, NJ 08544, USA.
Email: martincz@princeton.edu

Funding information
Waisman Center, Grant/Award Number: U54 HD090256; National Science Foundation, Grant/Award Number: GRFP DGE-1747503; National Institute of Child Health and Human Development, Grant/Award Number: F32HD110174 and R37HD037466

Abstract

How do infants become word meaning experts? This registered report investigated the structure of infants' early lexical representations by manipulating the typicality of exemplars from familiar animal categories. 14- to 18-month-old infants (N=84; 51 female; M=15.7 months; race/ethnicity: 64% White, 8% Asian, 2% Hispanic, 1% Black, and 23% multiple categories; participating 2022–2023) were tested on their ability to recognize typical and atypical category exemplars after hearing familiar basic-level category labels. Infants robustly recognized both typical (d=0.79, 95% CI [0.54, 1.03]) and atypical (d=0.70, 95% CI [0.46, 0.94]) exemplars, with no significant difference between typicality conditions (d=0.14, 95% CI [−0.08, 0.35]). These results support a broad-to-narrow account of infants' early word meanings. Implications for the role of experience in the development of lexical knowledge are discussed.

In their first years of life, infants rapidly learn about the meanings of words. As early as 6 months of age, infants recognize some frequently heard words as referring to familiar objects (Bergelson & Swingley, 2012; Kartushina & Mayor, 2019; Tincoff & Jusczyk, 1999). Infants' ability to link familiar words to possible referents becomes more refined with age and experience (Bergelson, 2020; Bergelson & Aslin, 2017); the number of words for which children can recognize referents increases rapidly across the second year of life (Braginsky et al., 2019; Frank et al., 2021). These word learning skills lay the foundation for children's early language development and predict a range of later linguistic, cognitive, and academic outcomes (Bleses et al., 2016; Marchman et al., 2018).

What is the nature of infants' early word meaning representations? As a starting point for this question, it is useful to consider the structure of word representations that will eventually emerge over development. Studies designed to uncover adults' knowledge of word meanings reveal rich underlying structure in the semantic associations within lexical networks (e.g., De Deyne et al., 2019; Steyvers & Tenenbaum, 2005) and in the potential referents to which word meanings extend (Goldberg, 2019). Rather than indexing a specific, narrow range of referents, each noun is connected to a rich category of interrelated referents. One fundamental organizing principle in adults' representations of word meanings is typicality. Adults extend the label *dog* to refer both to typical (e.g., beagles) and atypical (e.g., dachshunds) dogs, with typical exemplars and their features accessed more rapidly than atypical exemplars and features (Rosch, 1978). Infants will eventually become word meaning experts who can effortlessly apply the words they know to a broad range of category items—but when and how does this word knowledge emerge?

Abbreviation: OSF, Open Science Framework; MB-CDI, MacArthur-Bates Communicative Development Inventory.

[*]Haley Weaver and Martin Zettersten equal authorship contribution.

The development of word meaning representations—Broad-to-narrow or narrow-to-broad?

Two general accounts of infants' understanding of the extension of newly learned words have been advanced in past literature: a "broad-to-narrow" view and a "narrow-to-broad" view (Hirsh-Pasek et al., 2004). On the broad-to-narrow view, infants construe early-learned words as referring to broad categories of items, subsequently refining their understanding of a word's possible referents based on experience (Waxman & Gelman, 2009; Waxman & Markow, 1995; Yin & Csibra, 2015). This account predicts that infants use initial learning experiences to successfully link words to a wide range of both typical and atypical exemplars early on in development, supported by rich prior knowledge about category structure (Snedeker & Gleitman, 2004) or by general expectations that words will refer to categories (Ferguson & Waxman, 2017; Waxman & Markow, 1995). One source of evidence in support of this view is that children overextend their word meanings to referents that are perceptually similar, categorically related, or thematically associated with the target word referent (Gruendel, 1977; Huttenlocher & Smiley, 1987; Mayor & Plunkett, 2010; Rescorla, 1980). For instance, children may use the word *horse* to correctly refer to horses, but also include other semantically related animal kinds such as goats (Rescorla, 1980). The broad-to-narrow view suggests that as children amass experience with language and how it is used to refer to objects in their environments, they gradually refine their understanding of words' possible referents and become increasingly adept at distinguishing between closely related referents to fine-tune their word meanings. Thus, the broad-to-narrow account predicts that infants' broad early word meanings should have limited sensitivity to idiosyncratic experiences with individual category exemplars. Once infants grow older, individual differences in experience with possible word referents begin to play an increasingly important role in constraining word meanings.

By contrast, the narrow-to-broad view suggests that infants may first form restricted, exemplar-based representations of word meanings, and gradually begin to generalize words to encompass broader categories through extended experience with words co-occurring with a variety of exemplars (Ambridge, 2020; Barrett, 1986; Hirsh-Pasek et al., 2004; Smith et al., 2002). This account predicts that infants' early word meanings may be limited to familiar, typical exemplars. Indeed, some accounts of infants' lexical production suggest that they may initially use a label to refer to a specific object prior to correctly generalizing the label to the entire class of items (Gruendel, 1977; Rescorla, 1980). A second prediction is that infants' early word meaning representations should be closely linked to individual differences in experience, since the specific exemplars that infants encounter while hearing labels (e.g., hearing the word *dog* while seeing the family pet) will vary substantially and idiosyncratically across infants. In support of this hypothesis, there is evidence that infants extend early words that have highly familiar, salient referents more slowly than other kinds of words. For example, infants are slower to correctly generalize words such as *dog* or *Momma*, perhaps because infants' experience with these particular words is limited to exposure to a specific category referent (i.e., their own mother or pet dog; Rescorla, 1980). In sum, the narrow-to-broad view predicts that infants represent early word meanings as associated with a narrow set of category referents, and gradually incorporate a broader set of items into their word meaning representations through additional experience with words and their referents.

Prior research has provided empirical evidence that supports both accounts and has been limited in its ability to disambiguate these views for several reasons. First, most research investigating how infants generalize labels across category members introduces novel words and artificially constructed categories. While these studies have offered tremendous insight into the dimensions and features that infants use when generalizing word meanings, they leave unclear the nature of infants' early representations of words they encounter frequently in their day-to-day experience. Second, the vast majority of studies investigating familiar word knowledge focus on a limited set of prototypical category referents and rarely contrast infants' ability to link the same word with a range of exemplars (with notable exceptions that will be discussed below, e.g., Garrison et al., 2020; Meints et al., 1999; Perry & Saffran, 2017; Southgate & Meints, 2000). Third, mainly due to limitations in power in infant research, past work has often focused on whether infants exhibit word recognition *across* all participants and items, at the expense of interrogating individual infants' knowledge about specific items, limiting researchers' ability to assess the role of idiosyncratic experience in the development of word meanings. Next, we review past research investigating the development of infants' word meanings, focusing on (a) relationships between lexical and semantic networks, (b) individual differences in word meanings, and (c) past research on infants' understanding of words in relation to typical and atypical category exemplars.

Structure and variability in children's early representations of word meaning

Lexical-semantic networks

One window into the nature of infants' word meanings is the development of their early lexical networks. How infants represent the relationships between words in their budding vocabularies yields insight into how infants represent the underlying meanings of words. Past

research suggests that infants do not simply learn words as isolated representations; instead, they connect semantically related words from early on in development (Hills et al., 2009; Peters & Borovsky, 2019; Wojcik, 2018; Wojcik & Saffran, 2013). For example, by the age of two, infants distinguish associations between semantically related words, such as *dog* and *kitty*, from associations between semantically unrelated words (e.g., *dog* and *juice*) (Willits et al., 2013). The interconnected nature of infants' lexical-semantic networks has also been revealed through priming studies (Arias-Trejo & Plunkett, 2010; Mani et al., 2013). For example, infants will orient faster to a target image after hearing a familiar word (e.g., *cookies*) when it is preceded by a semantically related prime word (e.g., *bananas*) compared to an unrelated word (e.g., *house*). Thus, early vocabulary networks are organized around semantic properties of word referents. However, while studies on the development of early vocabulary networks elucidate the semantic relationships *between* words, they leave open the question of how broadly or narrowly infants represent the meanings of *individual* words.

Individual differences in word meanings

Children exhibit individual differences in their vocabulary networks, with consequences for how they represent known words and interpret new labels (Beckage et al., 2011; Colunga & Sims, 2017). For example, toddlers who know more terms for categories organized by material (e.g., *wood* or *chalk*) are more likely to generalize the meaning of a novel word based on material (Perry & Samuelson, 2011). The structure of infants' early vocabularies influences how easily they acquire novel words (Borovsky et al., 2016) and how readily they recognize familiar words in the presence of competing referents (Borovsky & Peters, 2019). These types of individual differences apply not only to children's spontaneous generalization of novel words but point to deeper individual differences in how infants represent the meanings of known words. In one study (Perry & Saffran, 2017), 21-month-olds were presented with familiar objects depicted in typical (e.g., a red strawberry) or atypical colors (e.g., a green strawberry). While infants correctly recognized words in the presence of both typical and atypical target referents, their word processing was significantly less accurate for objects with atypical colors, suggesting that by 21 months, infants are sensitive to prototypical features of familiar referents (like the color of a strawberry) during word recognition. Crucially, the degree to which infants experienced a processing disruption depended on their existing vocabulary structure: children who knew more words for object categories organized by shape were less disrupted, perhaps because they had a more strongly shaped-based representation of words such as strawberry. Word meanings likely vary substantially across infants, but this may not be evident when infants are tested only on prototypical category referents.

One recent study illustrates how individual differences in infants' early word meaning representations can be revealed by testing word recognition beyond standardized category exemplars (Garrison et al., 2020). Infants' word recognition for familiar as opposed to unfamiliar prototypical category referents was tested by asking caregivers to supply photographs of objects from infants' home environments. Infants, who ranged in age between 12 and 18 months, were then tested on how quickly and accurately they linked a known label (e.g., *book*) to the caregiver-supplied image of an object familiar to them (e.g., a book commonly read to them in the home) compared to an image of an unfamiliar object (e.g., an unfamiliar book). While older infants were similarly able to link both the familiar and the unfamiliar image to the target label, younger infants only showed above-chance word recognition for familiar items—consistent with the hypothesis that infants' early word meanings are associated with a narrow set of (potentially idiosyncratic) familiar exemplars. However, significant differences between familiarity conditions were inconsistent across analyses and no interaction between familiarity and age was observed, leading the authors to conclude that their data did not provide strong evidence for a narrow-to-broad change in word meanings.

Category member typicality in word recognition

Even young infants can form categories after experiencing perceptually diverse exemplars (Bornstein & Arterberry, 2010; Eimas & Quinn, 1994). Exemplar typicality plays a key role in early category learning; exemplar similarity and prototypicality ease infants' ability to form visual categories (Bauer et al., 1995; Oakes et al., 1997). Moreover, infants' individual experiences with exemplars from a given category influence their perceptual processing of category members (Hurley & Oakes, 2018; Kovack-Lesh et al., 2014) and the degree to which infants develop finer-grained sensitivity to distinctions within basic-level categories (e.g., distinguishing beagles from Saint Bernards within the dog category; Quinn & Tanaka, 2007). Given that infants' developing perceptual categories are organized by exemplar typicality, to what extent are infants' early representations of word meanings also constrained by category typicality?

There is some evidence to suggest that infants initially recognize referents that are prototypical category exemplars more easily than referents that are atypical exemplars when processing familiar words. Meints and colleagues (1999) tested infants' knowledge of known words—ascertained through parental report for each participant—using typical or atypical members of

the named category as referents in a looking-while-listening paradigm at 12, 18, and 24 months. Infants were presented with typical or atypical members of the named target category paired with typical or atypical members of a different category, and infants' fixation of the target image was measured after hearing the known label. 12-month-old infants increased their looking to typical category members (e.g., a sparrow) after word onset, but not to atypical category members (e.g., an ostrich). However, older infants (18- and 24-month-olds) recognized both typical and atypical category members in response to a category label. Relatedly, Poulin-Dubois and Sissons (2002) found that 18-month-old infants recognized incomplete images of category members in which some parts of the image were deleted (thus making these items, arguably, atypical exemplars) as instances of familiar words in a preferential looking task. These findings suggest that infants may initially have relatively narrow representations of known words, and subsequently form broader representations of a word's meaning.

However, limitations in the literature complicate this interpretation, leaving open the question of how word meaning extensions change across development. First, close inspection of Meints et al.'s (1999) analyses suggest a more modest conclusion regarding whether typical exemplars were recognized better than atypical exemplars by younger infants: while significant looking was found for typical exemplars and not for atypical exemplars, no significant difference in looking to typical versus atypical exemplars was reported for 12-month-olds, and no significant interaction showing change across age was observed. Similarly, Garrison et al. (2020) found suggestive but inconsistent evidence that object familiarity predicted word recognition. While younger infants showed successful word recognition only for familiar items, no consistent evidence supported a difference between familiar and unfamiliar items. Conversely, while some studies find similar patterns of word recognition for typical and atypical exemplars for infants from 18 to 24 months of age (Meints et al., 1999; Robinson, 2002), other results provide evidence for typicality effects among infants within this age window (Perry & Saffran, 2017; Southgate & Meints, 2000). For example, when presented with two exemplars from a given category (one typical and one atypical), 18- and 24-month-old infants preferred to look at the prototypical category member after hearing its label (Southgate & Meints, 2000). Thus, while some studies provide evidence in support of a narrow-to-broad developmental change in infants' early word meanings, questions remain both about whether infants' early word representations differ by typicality and the extent to which typicality effects on infants' word recognition disappear by 18 months.

A second open question in the literature is the role of experience in shaping early word meanings. A key prediction of the narrow-to-broad account is that infants' early word meaning representations should be principally associated with the specific exemplars that are highly representative of infants' idiosyncratic environments. However, there have been few direct tests of the degree to which infants' word representations vary across category exemplars as a function of experience. One limitation of past studies is that they have typically tested only one typical or atypical referent per category, which may obscure the extent to which infants differ in how they generalize a given word's meaning across a broader range of category members. While one study (Robinson, 2002) did test familiar word recognition for a wider range of category members in 17- to 25-month-olds, the results revealed no evidence of word recognition for either typical or atypical items among older infants, limiting the conclusions that can be drawn. A likely explanation of this absence of word recognition lies in the unusual variation of the preferential looking procedure used in the study, which included an 8-s baseline window and an 8-s looking window post-label onset, far longer than designs used to consistently elicit familiar word recognition (see e.g., Fernald et al., 2008). A second limitation of the current literature is that few studies have sought to directly quantify infants' experience in relation to their word representations (with some notable exceptions: Garrison et al., 2020; Robinson, 2002; Southgate & Meints, 2000). Southgate and Meints (2000) quantified object experience through caregiver report but did not relate this measure to infant behavior. Garrison et al. (2020) found no influence of the degree of similarity between familiar and unfamiliar items on infants' word recognition, but the typicality range of the items used in the study was limited due to the fact that even the unfamiliar items tended to be prototypical category members. Therefore, it currently remains unclear to what extent infants' early word meanings vary across individuals and whether such variation stems from differences in individual experience.

In sum, two key questions remain from past work: How narrow are infants' early word meanings, and to what degree do early word meanings vary with experience? First, while some past studies provide evidence that early word meanings are narrowly focused on typical items, the literature reveals inconsistencies that complicate this interpretation. Some authors argue for a narrow-to-broad developmental change such that there is a typicality advantage in early word meanings that dissipates by 18 months (Meints et al., 1999), while other studies suggest that early word meanings may not vary substantially according to experience typicality (Garrison et al., 2020) or that typicality effects in word recognition continue to be found beyond 18 months (Perry & Saffran, 2017). Second, while a narrow-to-broad account predicts that infants' early word representations should vary based on experience with specific category exemplars, we still have

little direct evidence relating experiential factors to variability in the breadth of infants' early word meanings. Conclusions about how broadly children generalize particular words across a given category are limited by the small number of category exemplars used in past studies and the limited number of attempts to relate infant word meanings directly to typicality of prior experience. In the current study, we seek to clarify these two questions by investigating developmental change in infants' word representations tested across a broader set of category exemplars, while using caregiver report to quantify exemplar typicality for each infant.

The present study

Are infants' early word representations initially limited to a narrow set of exemplars (e.g., that the word *dog* refers to a prototypical category member like Labrador) or do these representations include broad sets of exemplars from early in development (e.g., that *dog* may also refer to less representative category members such as chihuahuas, greyhounds, and pugs)? In the current study, we investigated the variability in infants' lexical representations across development by manipulating the typicality of exemplars from highly familiar categories. Specifically, we tested infants (14–18 months of age) on their ability to recognize typical and atypical category exemplars in a looking-while-listening experiment (Fernald et al., 2008). The words were a set of basic-level category nouns (*dog*, *bird*, *cat*, and *fish*) that most infants comprehend by 10–15 months of age (Supporting Information S2; Table S1; Frank et al., 2017). A small set of categories were used to ensure that we could include a wider range of exemplars per category, varying in typicality. Infants were tested in two sessions in order to maximize the number of observations per infant, and caregivers completed a questionnaire on their infants' experiences with each category. The study protocol was peer-reviewed and preregistered as a registered report (Stage 1 manuscript: https://osf.io/rqbtn). All reported methods and analyses were preregistered unless specified otherwise. Our study pursued three main aims.

Aim 1: To determine whether the typicality of category exemplars affects infants' lexical processing

Specifically, is lexical processing more robust for referents that are typical category exemplars than referents that are atypical category exemplars? That is, are infants' word meanings initially narrow (faster and more accurate lexical processing for typical category exemplars) or broad (similar processing for typical and atypical category exemplars)?

Aim 2: To investigate how typicality effects change with age

A narrow-to-broad account would predict that the effect of category typicality on word recognition interacts with age, such that only more typical category referents are recognized reliably by younger infants, and the successful recognition of atypical category referents emerges with age. A broad-to-narrow account would predict early success at recognizing both typical and atypical category referents from an early age, with typicality differences emerging later in development as infants accrue additional experience with words and their possible referents (i.e., no interaction between typicality and age given the young age of infants in our current sample, or an interaction such that typicality differences emerge late).

Aim 3: To test whether individual differences between participants in word recognition and typicality effects are predicted by differences in experience (via parental report)

Our third aim was to investigate whether individual differences in the effects of typicality were related to differences in infants' caregiver-reported experiences with the specific categories and exemplars tested in this experiment. A narrow-to-broad account of word meanings would predict that variability in children's early experience with category referents explains, and therefore should predict, the magnitude of typicality effects on word recognition, such that word recognition is more robust for an exemplar that is typical *for that child*. On a broad-to-narrow view, words are predicted to be extended to a broad class of category members even at a young age, such that word meanings should be less sensitive to early experience with individual exemplars. This account thus predicts little to no overall effect of individual variation in early experience on word recognition for infants in our sample.[1]

METHODS

Participants

Families were recruited from the United States through Lookit (Scott & Schulz, 2017; https://lookit.mit.edu/), an online infant testing platform. The study received IRB approval to be conducted online using the Lookit

[1]As in Aim 2, a broad-to-narrow account predicts that experience gradually helps learners refine and constrain initially broad word meanings, such that individual differences in experience may play an increasing role in word recognition as children grow older. However, in Aim 3, our primary focus is on the overall effect of individual differences in exemplar experience on word recognition in our current sample (rather than its interaction with age).

platform. Parents signed up to be included in a participant database and were notified of their child's eligibility for studies via email. English-learning infants between 14 and 18 months of age ($N=84$; 51 female infants; $M=478$ days; range: 410–574) that had no history of developmental concerns participated in two sessions approximately 1 week apart. Caregiver-reported race/ethnicity for the infants was 64% White, 8% Asian, 2% Hispanic, 1% African American, and 23% reporting multiple categories. Families received gift cards for participating in the study.

The sample size was determined by conducting a power analysis using the jpower module in jamovi (The jamovi project, 2020) and considering resource constraints (Lakens, 2022). Our registered sample size of $N=80$ provided 90% power to detect an effect size of $d=0.37$ (and 80% power to detect an effect size of $d=0.32$) in the main test of the typicality effect (Aim 1). The target sample size reflected the goal of striking a balance between having adequate power to detect effects large enough to be of theoretical interest and our data collection constraints for the project, based on the resources required for recruitment, compensation for participants, and video coding. In the Stage 1 registered report, we estimated the maximum number of participants we could recruit and test in the current experiment to be approximately 100 families. Assuming an estimated dropout rate of 20%, we arrived at the target sample size of $N=80$ and determined that this sample provided sufficient power to detect effect sizes within a range that is of most theoretical and practical significance (see "Analysis plan" section below and Supporting Information S8 for an extended discussion of power with respect to the analyses of interest). A detailed list of exclusionary criteria at the participant level can be found in the Supporting Information (see S1.1 Participant-Level Exclusions). We planned to replace infants that did not meet eligibility criteria until a final sample of 80 participants was reached.

Ultimately, we tested 143 children (substantially more than initially planned), resulting in a total of 255 individual testing sessions, after filtering out participation attempts beyond a given participant's first two completed sessions. There was a 22% attrition rate ($N=31$) from session 1 to session 2. 27 of these 31 participants without session 2 data (19% of all participants) subsequently did not provide sufficient data to meet our inclusion criteria (at least 24 valid trials) and were excluded. An additional 32 participants (22%) were excluded due to insufficient data contributions after excluding trials with poor video quality, technical errors, low frame rates, and/or parental interference (see S1.1 and S1.2 for further details). Eye-gaze data collected via Lookit must be manually coded prior to applying exclusionary criteria. We therefore recruited and coded infants in batches until we reached our final sample of eligible participants. This process resulted in a slightly larger sample size compared to the planned sample size in the Stage 1 registered report ($N=84$ vs. the registered $N=80$). A detailed explanation can be found in the deviations from the S1 registered report section below.

We chose the current age range (14–18 months) based on three main considerations. First, previous research suggests that there may be developmental changes in infants' word recognition for category exemplars varying in typicality and familiarity between 12 and 18 months of age (Garrison et al., 2020; Meints et al., 1999). Second, recent research suggests that there is a qualitative shift in infants' performance on looking-while-listening tasks at approximately 14 months (Bergelson, 2020). Given that we plan to evaluate individual differences in word recognition, we chose an age range during which infants' word recognition in this paradigm becomes more robust. Third, in our pilot data, we observed that infants below 14 months of age showed less robust word recognition than infants over 14 months of age, consistent with the quantitative shift observed in previous research (see S7.2 Online Pilot Experiment Results in the Supporting Information for further details).

Stimuli

All stimuli and experimental materials are available on the Open Science Framework (OSF) at https://osf.io/3t8gf/.

Animal words/categories

We selected four basic-level animal categories for which to test infants' word recognition. The superordinate category of animals was selected because we expected that the items would elicit high engagement across this age range. The category labels were a set of words (*dog*, *bird*, *cat*, and *fish*) selected according to the following criteria: (a) All words are understood at a relatively young age (see S2.1 Word Properties in the Supporting Information) and (b) exemplars from the corresponding basic-level category vary substantially in terms of typicality. All words occur frequently in corpora of child-directed speech in American English (see Supporting Information S2.1).

Images

The stimuli were a subset of 70 animal images collected from open-source image databases (Brodeur et al., 2010; Emberson & Rubinstein, 2016; Wikimedia Commons; https://unsplash.com). Images were edited to isolate the target animal and placed on a white

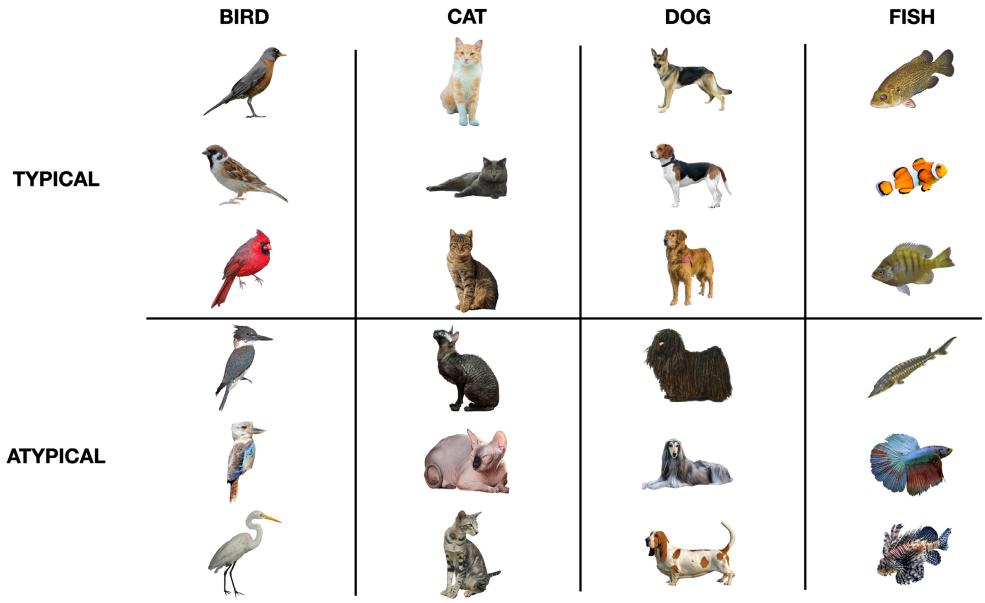

FIGURE 1 Typical and atypical items for each category.

600 × 600 pixel background. For each category (*dog*, *cat*, *bird*, and *fish*), we selected 16–20 images that varied in typicality. We then conducted a norming study with adults to collect objective ratings of typicality and nameability for all items within a category (see Supporting Information S2.2 and S3 for further details). The final set of items included three typical and three atypical exemplars from each of the animal categories ($n = 24$) that were likely to be named by adults using the basic-level category label (see Figure 1; S2.2 in the Supporting Information provides details on typicality ratings). Three images of naturalistic scenes (e.g., mountains) were selected to serve as attention getters.

Audio

A female native speaker of American English recorded 16 target sentences in child-directed speech. The sentences included four different carrier phrases: "Look at the [target label]", "Find the [target label]", "Do you see the [target label]", and "Where's the [target label]". Sentences were recorded in a sound-attenuated booth using Praat (Boersma & Weenink, 2019). All sentences were normalized to a total length of 1650 ms and an average intensity of 65 dB. 1500 ms of silence prior to the onset of the carrier phrase and 4000 ms of silence after the offset of the target word were appended to each recording, such that each auditory stimulus had a total duration of 7150 ms. Sentences were edited such that onset of the target label was consistent across all recordings (always 2650 ms after the onset of the auditory stimulus). Attention getters were presented with 5200 ms of instrumental music.

Design and procedure

Families were alerted of their infants' eligibility via the Lookit platform and provided video documentation of consent. Before beginning the experiment, parents were asked to complete an at-home set-up to maximize video recording quality by reducing backlighting and home distractions. Parents were asked to seat their child comfortably in front of a laptop or desktop computer and close their eyes or turn around to limit their influence on infants' looking behavior.

Infants participated in two data collection sessions approximately 1 week apart ($M = 8$ days, Mode: 6 days, range: 0–104 days). Word recognition for the four basic-level animal categories was assessed using a looking-while-listening paradigm presented on Lookit (Scott & Schulz, 2017). During each session, infants were presented with 24 trials with animal referents that varied in typicality. Attention getters were distributed evenly throughout the experiment to maintain interest in the task.

On each trial, infants saw two animal images that were matched on typicality and heard a phrase labeling one of the animals. For example, a typical trial for the target noun *bird* presented an image of a robin and an image of a tabby cat, while an atypical trial presented an image of a kingfisher and an image of a sheepdog (see Figure 2). Pairs of items from different basic-level categories were yoked together (e.g., sparrow always appeared with beagle) such that each basic-level category occurred equally often with each of the other basic-level categories. Each item appeared twice as the target and twice as the distractor for a total of four exposures across two sessions.

Trials were presented in six blocks of four separated by an attention getter, for a total of 24 trials per session.

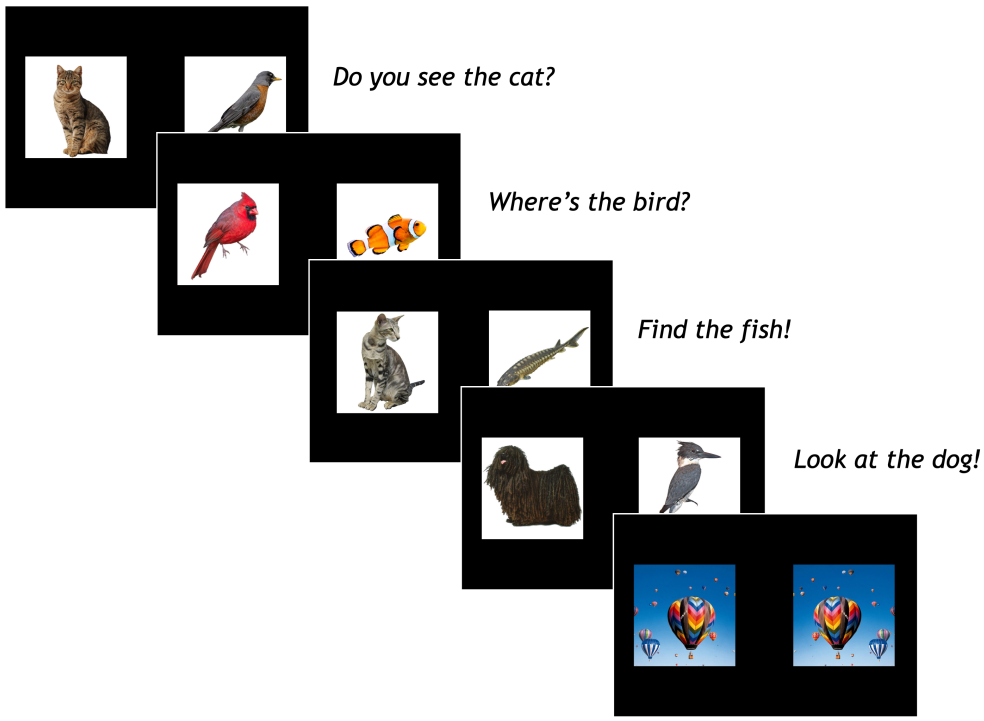

FIGURE 2 Sample trial order for one block. Each target item was tested once per block with half of the trials displaying typical items and half displaying atypical items.

Within a block, infants saw two typical and two atypical trials testing each of the four target labels (*bird*, *cat*, *dog*, *fish*) displayed in a pseudorandom order so the same target item did not appear more than twice in a row (Figure 2). The order of trial presentation and the target location was counterbalanced across participants and testing sessions. Within a single session, each basic-level category label was tested six times and the location of the target item appeared equally often on the left and right side of the screen. For the second session, both the order of trial presentation and target location of the category label were reversed (see the OSF page for sample trial orders).

Parental survey

Parents completed a survey after their infant completed the second testing session. In our Stage 1 manuscript, we planned that the survey would be composed of two components—a measure of children's vocabulary size (MB-CDI (MacArthur-Bates Communicative Development Inventory) Level I Short Form, slightly modified to include all four category labels tested in the study; Fenson et al., 2000) and a survey on infants' prior experience with each category exemplar presented during the experiment. The final survey deviated from the Stage 1 registered report proposed survey in that we did not collect the measure of children's vocabulary size and only included questions about infants' prior experience with each category exemplar. Parents judged each of the 24 items presented during the experiment in terms of typicality for their child (*experience-based typicality*). For each image, parents were asked "How typical is this image of the [category label] your child experiences?" and to rate the image on a scale from 1 (very atypical) to 5 (very typical). Parents were also asked to report the contexts in which their infant experiences each animal (e.g., as a pet at home, in a book you read, at the local zoo, etc.). The full instructions for the survey can be found in the Supporting Information (see S4).

Video coding

Infants' eye-gaze behavior was coded frame-by-frame from webcam video recordings. Details on our procedure for coding videos and assessing reliability is reported in the Supporting Information (S5). Reliability between coders was high (93.9% average frame agreement).

Pilot data

We conducted a pilot experiment to test the feasibility of our design. In the Supporting Information, we include details on the experiment methods (S7.1) and preliminary results demonstrating that we can successfully measure word recognition on Lookit (S7.2).

Deviations from the S1 registered report

Larger sample size

As noted above, the experiment included a larger sample size than originally registered ($N=84$ as opposed to $N=80$ in the Stage 1 manuscript). Given the asynchronous nature of the Lookit eye-gaze data, we were unable to replace ineligible infants one-by-one. Instead, we collected and coded data (from July 2022 to September 2023) in batches until we had at least 80 eligible participants, resulting in a final sample of $N=84$. Critically, no data-dependent decisions were made with respect to sample size (beyond evaluating whether infants met inclusion criteria). We also report an analysis restricting the sample to our original target of $N=80$ in Supporting Information S10.4, in which we find essentially equivalent results to our main analyses using the larger sample.

Wider age range

We included infants from a slightly broader age range (410–574 days) than we preregistered in our Stage 1 manuscript (410–560 days). Lookit recruitment did not allow for fine-grained restrictions on participant age, and we therefore also collected data from infants several days older than our preregistered age range. We expanded our age range because (a) we saw no principled reason to exclude infants who were only a few days older than our target age range, (b) our main analyses evaluated age in a continuous manner (Aims 2 and 3), and (c) including a wider age range (if anything) increased the power of these analyses.

MB-CDI data

We did not include the MB-CDI Level I Short Form or parent judgments of their infants' knowledge of the four category labels in the caregiver survey. This was an oversight on the part of the authors. Consequently, we were unable to conduct one preregistered robustness analysis (4.2 in the Stage 1 manuscript; see S9.1.2 in the Supporting Information) testing the condition effect after removing words that parents reported as unknown by their infants.

Video coding

We made one minor deviation from the planned video coding procedure, opting to retain all primary coding files for analyses (see S5.2 in the Supporting Information for details).

Overview of analytic approach

Modeling framework

All data processing and analysis was conducted using software packages in R (R Development Core Team, 2023; version 4.3.2). Linear mixed-effects models were fit using the lme4 package (Bates et al., 2015); *p*-values for mixed-effects models were computed using the lmerTest package (Kuznetsova et al., 2017), which uses the Satterthwaite approximation to estimate degrees of freedom for statistical tests. Predictor variables were centered unless otherwise specified. For each mixed-effects model specified below, we first fit a model with the maximal random-effects structure at the participant level (Barr et al., 2013). In models fit to trial-level looking data, we also included random intercepts for target word to account for non-independence due to items. In cases where the models with the specified random-effects structure failed to converge, we iteratively pruned random effects until convergence was achieved, removing random effects of lesser theoretical interest prior to removing random effects of greater theoretical interest. Specifically, we first removed random effects for covariates, then random effects for covariances among random effects (beginning with covariances close to zero), followed by random intercepts for target word and then for participant (Brauer & Curtin, 2018). Random slopes for the effects of interest (typicality condition and experience-based typicality) were removed only if all other pruning efforts were unsuccessful.

Windows of analysis

In our main analyses, we investigated the difference between children's proportion looking to the target image during a *critical window* after the onset of the target word and during a *baseline window* prior to the onset of the target word. The baseline window was set to encompass the 2 s immediately preceding the target word onset (−2000 to 0 ms), consistent with past literature (e.g., Garrison et al., 2020). The size of the critical window varies across the word recognition literature, with some experiments analyzing shorter windows of 300–1800 ms or 367–2000 ms post target word onset (Fernald et al., 2008; Garrison et al., 2020; Swingley & Aslin, 2002) and other experiments analyzing longer time windows extending as long as 3500 ms post target word onset (Bergelson & Swingley, 2012; Meints et al., 1999). The choice in window selection carries a tradeoff between limiting the focus of the analysis on infant behavior that is most closely related to the onset of the target word (in the case of the smaller windows of analysis), on the one hand, and capturing as much relevant looking behavior as possible, on the other. In particular, recent research suggests that longer time windows may increase reliability, a key concern in estimating individual differences (Zettersten et al., 2021). Given that we tested a wide age range in the current study, we set our critical window to lie between 300 and 2800 ms, as a compromise between these tradeoffs. We also investigated the degree to which the results depended on the length of the critical window (see S9.1.1 and S9.2.1).

Trial-level data processing and exclusion

Only trials with sufficient looking data (>50% looking during both the critical and baseline window) and without procedural errors were included in analyses. Trial-level exclusion criteria can be found in the Supporting Information (S1.2).

Dependent measure: Baseline-corrected proportion target looking

We computed the proportion of looking at the target item as opposed to the distractor item for both the baseline window (−2000 to 0 ms) and the critical window (300–2800 ms): target/(target + distractor). We then computed a baseline-corrected proportion target looking measure, by subtracting proportion looking to the target during the baseline window from proportion looking to the target during the critical window for each trial. Baseline correction is used to account for potential saliency differences between target and distractor images (e.g., Garrison et al., 2020; Meints et al., 1999). For models testing the average typicality effect at the participant level (Aim 1, "Average participant-level analysis" section), we computed the average baseline-corrected proportion target looking for both typical and atypical items. For models fit to trial-level data (i.e., the model in the "Trial-level analysis" section and the models in Aims 2 and 3), models were fit to baseline-corrected proportion target looking for each individual trial.

Stage 1 registered analysis plan

Aim 1: To determine whether the typicality of category exemplars affects infants' lexical processing

Average participant-level analysis

To investigate whether there was an effect of typicality on infants' word recognition, we first conducted an overall analysis averaging across all items and categories that participants encountered during the study. For each infant, we computed the average baseline-corrected proportion of looking to the target within the critical window across trials for each typicality condition (i.e., each infant provided two baseline-corrected proportion looking scores: an average baseline-corrected target looking score for typical items and an average baseline-corrected target looking score for atypical items). We report the mean and 95% CI of baseline-corrected target looking for each typicality condition as descriptive statistics. To test our questions of interest, we fit a linear mixed-effects model predicting average baseline-corrected proportion target looking from typicality condition (centered), including a by-participant random intercept and random slope for typicality condition. The model was specified as follows:

average_corrected_target_looking ~ 1 + typicality_condition + (1 + typicality_condition|participant)

Using this model, we answered the following three questions:

Do infants successfully recognize the target word? (quality check). Chance-level baseline-corrected proportion target looking is 0. The intercept of the model therefore tested whether infants successfully recognized the target words (averaging across typicality conditions, because typicality condition was centered). Testing whether infants showed successful word recognition overall served as a quality check of our method, ensuring that infants were engaging in the task as expected. We treated infants successfully recognizing the target words on average as a basis for interpreting effects of typicality: typicality effects would only be interpreted as meaningful if infants showed above-chance word recognition collapsing across typicality conditions. Our target sample size (N=80) ensured that we had over 99% power to detect an effect of d=0.5 or larger. An effect size of d=0.5 is smaller than the effect typically found in studies of online familiar word recognition (meta-analytic effect size: d=1.24; Bergmann et al., 2018), including for infants between 14 and 18 months of age (e.g., Garrison et al., 2020: d ~0.5–1 across an age range of 12–18 months).

Do infants differentially recognize typical versus atypical items? The main effect of typicality is statistically equivalent to a paired-samples t-test and tested whether there were differences in infants' recognition of typical items compared to atypical items. Given our sample size of N=80 participants, we had 90% power to detect an effect size of d=0.37 (80% power to detect an effect size of d=0.32). Past studies most similar to the current design have reported a wide spectrum of absolute effect sizes ranging from d=0.03 to d=0.64 (see Supporting Information S6 for an overview; Garrison et al., 2020; Meints et al., 1999; Perry & Saffran, 2017; Southgate & Meints, 2000). Our design was well-powered to detect an effect consistent with the upper range of effect sizes observed in past studies, but had limited power to detect small effect sizes in the lower range (see below for our rationale in choosing the smallest effect size of interest). A significant effect in this analysis would lead us to conclude that the typicality of category exemplars made a difference in how easily infants recognize familiar words. Specifically, we predicted that infants would show a higher average baseline-corrected proportion looking to the target item when it was typical than when it was atypical.

A non-significant effect in this analysis would not allow us to infer that there was no effect of typicality on infants' looking behavior. Instead, we used equivalence testing to assess whether the effect was smaller than a meaningful

effect size threshold, sometimes called the smallest effect size of interest (Lakens et al., 2018). We selected our smallest effect size of interest using the small-telescopes approach (Simonsohn, 2015), according to which the lower bound for equivalence testing is set as the effect size earlier studies would have had 33% power to detect. The most relevant studies from past literature that we identified have used sample sizes of approximately 30–40 participants (Garrison et al., 2020; Meints et al., 1999; Perry & Saffran, 2017). The smallest effect size that studies with this sample size would have 33% power to detect is approximately d=0.25, which we chose as our lower threshold for an effect size of interest. If the typicality effect was not significant, we would conduct an equivalence test evaluating whether the null hypothesis that the absolute effect size was at least as large as d=0.25 could be rejected, and the alternative hypothesis that the effect size lay between −0.25 and 0.25 was accepted. In other words, this test allowed us to conclude whether a non-significant effect was sufficiently small to be considered practically equivalent to a null effect.

Do infants successfully recognize words in the typical condition and in the atypical condition? We further tested whether infants successfully recognized words in each typicality condition within the same model, by recoding the typicality condition predictor in the model to be centered on the typical condition (atypical=−1; typical=0) and on the atypical condition (atypical=0; typical=1). For each of these re-centered models, the intercept now indicated whether infants' baseline-corrected looking during the critical window was significantly above chance for each of the typicality conditions. This design provided us with at least 80% power to detect an effect size of d=0.32 or larger. As in the analysis above, if either effect was non-significant, we would conduct an equivalence test with a threshold of d=0.25 to assess whether the effect size was statistically equivalent to an effect of minimal interest.

Trial-level analysis

To further test the effect of typicality, we also fit a linear mixed-effects model predicting trial-by-trial baseline-corrected proportion target looking from typicality condition (centered), including a by-participant random intercept and random slope for typicality condition as well as a random intercept for target word (i.e., *dog*, *cat*, etc.). Modeling trial-level looking data while accounting for non-independence due to the target word allowed us to assess the degree to which typicality effects generalized across particular category labels. The model was specified as follows:

$$\text{corrected\_target\_looking} \sim 1 + \text{typicality\_condition} + (1 + \text{typicality\_condition}|\text{participant}) + (1|\text{target\_word})$$

The a-priori power analyses reported in the Stage 1 manuscript and our plan for interpreting potential discrepancies between modeling results is now included in the Supporting Information (S8.1).

Timecourse analysis

While the previous analyses tested overall differences in infants' accuracy in recognizing typical and atypical category exemplars, they offer limited insight into how infants' processing unfolded across time. In order to gain finer-grained insight into the timecourse of infants' word recognition for typical compared to atypical items, we conducted a cluster-based permutation analysis (Maris & Oostenveld, 2007). This analysis identified where over the course of a test trial looking to the target diverged between the two typicality conditions, while controlling Type I error rate. The cluster-based permutation analysis was conducted using the eyetrackingR package (Dink & Ferguson, 2015). We report the full details of our approach from the Stage 1 manuscript in Supporting Information (S8.2).

Aim 2: To investigate how typicality effects change with age

To investigate whether the overall effect of typicality depended on participant age, we extended the trial-level linear mixed-effects model from 1.2. to include an interaction with participant age. We fit a linear mixed-effects model predicting participants' trial-by-trial baseline-corrected proportion target looking from typicality condition, participant age (mean-centered) and their interaction, including by-participant and by-target word random intercepts and a by-participant random slope for typicality condition. A significant interaction between typicality condition and age such that the typicality effect decreased with age would provide evidence consistent with a narrow-to-broad account of early word representations, suggesting that word representations are initially focused on a narrow set of category exemplars and increase in generality with age. Non-significant interactions (or interactions in the opposite direction, for example, increasing effects of typicality with age) would be interpreted as evidence against a narrow-to-broad account of early word representations. The model was specified as follows:

$$\text{corrected\_target\_looking} \sim 1 + \text{typicality\_condition} \times \text{age} + (1 + \text{typicality\_condition}|\text{participant}) + (1|\text{target\_word})$$

We report the a-priori power analyses from the Stage 1 manuscript in Supporting Information (S8.3).

Aim 3: To test whether individual differences between participants in word recognition and typicality effects are predicted by differences in experience

To address the third aim, we investigated whether individual differences in children's word recognition was related to caregiver report of infants' typical experience with specific category exemplars, while controlling for

age. In this analysis, the focus was on testing whether experience-based typicality was related to the typicality effect on participant accuracy. We fit a linear mixed-effects model predicting participants' trial-by-trial baseline-corrected proportion target looking from experience-based typicality (for a given target image/category exemplar) and age (mean-centered). We included random intercepts for participant and target word as well as a by-participant random slope for experience-based typicality. The item-level experience-based typicality measure was *z*-scored within a given participant. The model was specified as follows:

$$\text{corrected\_target\_looking} \sim 1 + \text{item\_experience\_typicality} + \text{age} + (1 + \text{item\_experience\_typicality}|\text{participant}) + (1|\text{target\_word})$$

The power analysis from the Stage 1 manuscript is reported in Supporting Information (S8.4).

Robustness analyses

We also conducted a series of robustness analyses to probe the degree to which any results hinged on key analytic decisions. To view the analysis plan of the robustness analyses preregistered in the Stage 1 manuscript, see Supporting Information S9.1. All robustness analyses are reported in S9.2, with the exception of the planned analysis excluding items parents reported as unknown on the MB-CDI, because we deviated from our Stage 1 plan and did not collect MB-CDI data.

RESULTS

All data and analytic scripts are openly available on the OSF at https://osf.io/3t8gf/. A walkthrough of all preregistered and exploratory analyses is available at https://rpubs.com/zcm/categories.

Aim 1: Typicality and lexical processing

Average participant-level analysis

Do infants successfully recognize the target word? (quality check)

Infants successfully recognized the target words (Model Intercept: $b = 0.07$, 95% CI [0.06, 0.09], $t(83) = 8.38$, $p < .001$). Averaging across participants, infants' proportion looking to the target increased by 0.072 (95% CI [0.055, 0.089]) relative to baseline, significantly above chance (chance = 0).

Do infants differentially recognize typical versus atypical items?

There was no significant effect of typicality, $b = 0.02$, 95% CI [−0.01, 0.04], $t(83) = 1.26$, $p = .21$, Cohen's $d = 0.14$ [−0.08, 0.35] (Figure 3). The equivalence test was also not statistically significant, $t(83) = -1.03$, $p = .15$. We therefore could not reject the null hypothesis that the absolute effect size was at least as large as $d = 0.25$.

Do infants successfully recognize words in the typical condition and in the atypical condition?

Infants robustly recognized the target words for both typical (Model: $b = 0.08$, 95% CI [0.06, 0.10], $t(83) = 7.24$, $p < .001$; Cohen's $d = 0.79$ [0.54, 1.03]; baseline-corrected looking: $M = 0.080$, 95% CI [0.058, 0.103]) and atypical exemplars (Model: $b = 0.07$, 95% CI [0.05, 0.08], $t(83) = 6.46$, $p < .001$, Cohen's $d = 0.70$ [0.46, 0.94]; baseline-corrected looking: $M = 0.065$, 95% CI [0.045, 0.085]).

Trial-level analysis

The model with the maximal random-effects structure yielded a singular fit that was only remedied by removing the by-participant random slope for typicality condition. However, the (singular) model including the typicality random slope yielded virtually identical results to the converging model including random intercepts for participant and target word only. As in the average participant-level analysis, infants' overall recognition of target words was significant in the trial-level model ($b = 0.07$, 95% CI [0.05, 0.10], $t(5.9) = 6.52$, $p < .001$) and there was no significant effect of typicality ($b = 0.01$, 95% CI [−0.01, 0.04], $t(3014) = 0.92$, $p = .36$). Word recognition was robust both for typical ($b = 0.08$, 95% CI [0.05, 0.10], $t(10.1) = 6.16$, $p < .001$) and atypical exemplars ($b = 0.07$, 95% CI [0.04, 0.09], $t(10.1) = 5.27$, $p < .001$).

Timecourse analysis

In the cluster-based permutation analyses, we found one cluster of adjacent time bins ranging from 0 to 200 ms with $|t| > 2$ (in the direction of higher accuracy for typical exemplars compared with atypical exemplars). However, this cluster did not reach significance in the permutation test, $p = .41$.

Aim 2: Effects of typicality across age

In the trial-level linear mixed-effects model including age, typicality condition, and their interaction, we found a significant effect of age ($b = 0.02$, 95% CI [0.01, 0.03], $t(88.6) = 3.33$, $p = .001$), suggesting that word recognition accuracy increased with age overall. There was no significant interaction between age and typicality ($b = 0.004$, 95% CI [−0.01, 0.02], $t(3015) = 0.52$, $p = .60$), meaning that we found no evidence that the effect of typicality changed with age (Figure 3b).

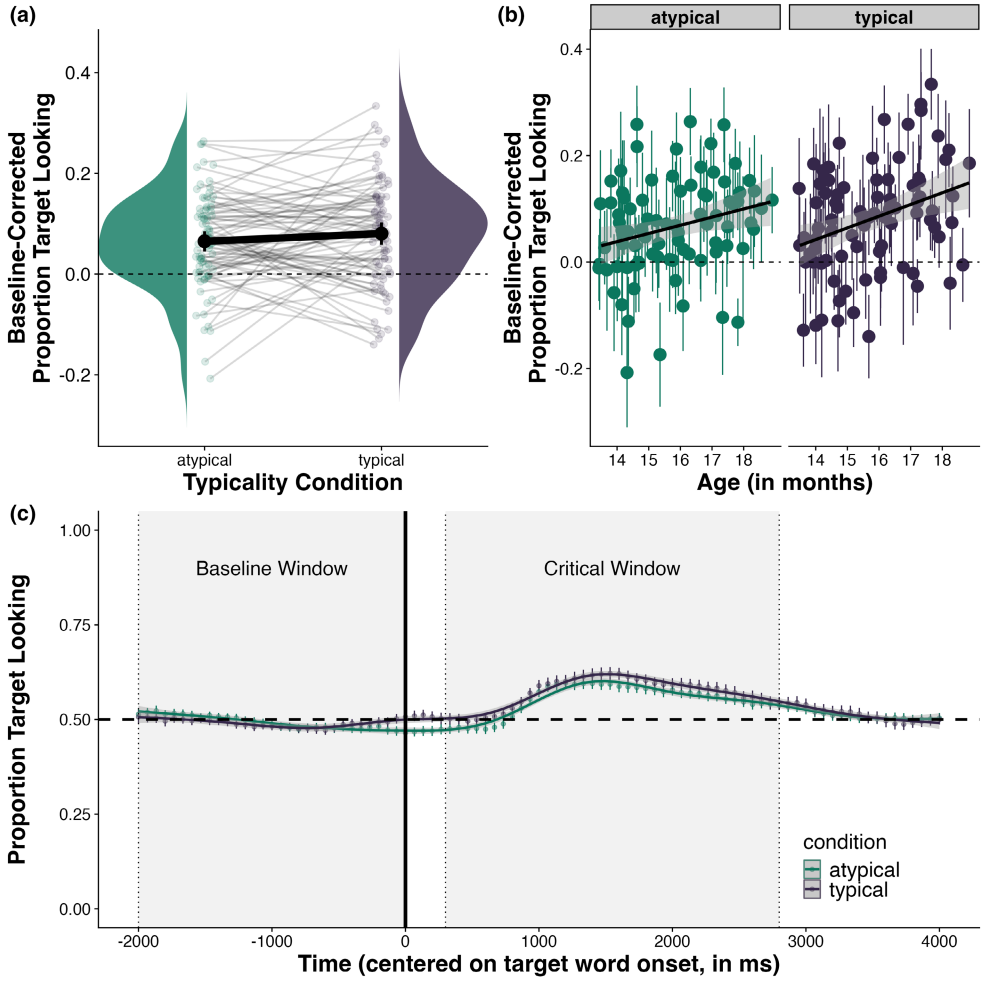

FIGURE 3 Word recognition depending on typicality condition. (a) Average baseline-corrected proportion target looking for each condition (in black). Individual points represent individual subjects, lines link subject responses between conditions. Error bars represent 95% CIs. (b) Baseline-corrected target looking for each condition by age. Each point represents the average for an individual subject, with 95% error bars. The regression line represents a linear fit with 95% error bands. (c) Timecourse of average proportion target looking for typical and atypical items. Error bars are ±1 SEs. Smoothed fits are based on a general additive model using cubic splines to visualize average looking trajectory.

Aim 3: Predicting lexical processing from individual differences in experience

Only infants whose caregiver completed the parent-report survey were included in the linear mixed-effects model testing individuals' experience-based typicality, leading to a reduced sample relative to the previous analyses ($N = 74$). The main preregistered model including a by-participant random slope for experience-based typicality yielded a singular fit. Attempts to prune other random effects were unsuccessful, so we removed the random slope for experience-based typicality. However, the model estimates for the singular fit model were very similar to the estimates for the final converging model omitting this random effect. Caregiver report of exemplar typicality did not significantly predict infants' baseline-corrected word recognition accuracy ($b = 0.004$, 95% CI [−0.01, 0.017], $t(2255) = 0.56$, $p = .58$). Controlling for parent-reported typicality, age remained a significant predictor of infants' word recognition ($b = 0.015$, 95% CI [0.003, 0.027], $t(67) = 2.47$, $p = .016$).

Robustness analyses

We report the results from all robustness analyses preregistered in the Stage 1 manuscript in Supporting Information S9.2. There were three key results: (1) we obtained qualitatively equivalent results (e.g., the same patterns of significance) when using a shorter critical window; (2) we observed no typicality effects in reaction time; and (3) there were no interaction effects with test session.

Exploratory analyses

Below, we report several exploratory results. Additional exploratory analyses are reported in Supporting Information (S10), including an assessment of

TABLE 1 Overview of average baseline-corrected proportion target looking for each target word.

Target word	Typical exemplars	Atypical exemplars	Typicality effect
Bird	M=0.075 95% CI [0.038, 0.111]	M=0.091 95% CI [0.048, 0.134]	$t(83)=-0.66$, $p=.51$
Cat	M=0.103 95% CI [0.063, 0.143]	M=0.075 95% CI [0.037, 0.114]	$t(83)=1.07$, $p=.29$
Dog	M=0.073 95% CI [0.038, 0.109]	M=0.025 95% CI [−0.014, 0.065]	$t(83)=1.77$, $p=.08$
Fish	M=0.064 95% CI [0.022, 0.105]	M=0.068 95% CI [0.025, 0.110]	$t(83)=-0.13$, $p=.89$

measurement reliability (S10.2), a sensitivity analysis testing the main typicality effect across a range of trial-based inclusion criteria (S10.3), a robustness analysis testing the main models when restricting the sample to our originally planned N of 80 participants (S10.4), a robustness analysis investigating alternative model specifications (S10.5), and an analysis predicting proportion target looking from the visual similarity of target and distractor images (S10.6).

Is there evidence for a typicality effect using continuous adult typicality norms?

The images used in the current study were rated on a Likert scale from very atypical (1) to very typical (5) as part of the adult norming study (Supporting Information S3). We used these ratings to categorize exemplars in a binary fashion as either typical or atypical. However, typicality is a continuous feature of category exemplars, with some category members being more or less typical of the category relative to others. In an exploratory analysis, we therefore examined the effect of typicality when using the continuous typicality norms. We fit the trial-level linear mixed-effects model from Aim 1, replacing the binary typicality predictor with the z-scored adult typicality ratings. We again did not find a significant effect of typicality, b=0.012, 95% CI [−0.001, 0.024], $t(2842)=1.82$, $p=.07$, though note that the p-value was marginal.

Item-level analysis

We also explored the degree to which word recognition varied across items, both on the level of target words and target images.

Target word

Broadly speaking, we found robust word recognition across all four target words (bird: M=0.08, 95% CI [0.05, 0.11]; cat: M=0.09, 95% CI [0.06, 0.12]; dog: M=0.05, 95% CI [0.02, 0.07]; fish: M=0.07, 95% CI [0.04, 0.10]). Average baseline-corrected looking to the target was similar within each typicality condition (Table 1; see S10.1 in the Supporting Information for a graph showing by-target-word variation across participants). The category label *dog* showed the strongest evidence of a typicality effect numerically, but even here there was no significant difference between typical and atypical dog exemplars ($p=.08$).

Target image

While there was wide variation in average baseline-corrected proportion target looking across the 24 image exemplars, word recognition was robust across items (Figure 4). In particular, 23 of 24 baseline-corrected mean estimates for individual target items were above 0, and the 95% confidence intervals for 15 out of the 24 items excluded 0.

Alternative model specifications

In exploratory analyses, we also considered alternative methods of specifying linear mixed-effects models incorporating information about target looking during the baseline window and the critical window. Here, we report the results from models predicting target looking during the critical window while controlling for baseline looking. In supplementary analyses (S10.5), we also consider a model in which we predict target looking from the interaction between typicality and trial window (critical window vs. baseline window).

Predicting proportion target looking during the critical window while controlling for baseline looking

An alternative to baseline correction is to predict (uncorrected) trial-level proportion target looking during the critical window while controlling for proportion target looking during the baseline window. This method may in principle provide more power, because it allows the relationship between looking during the baseline window and the critical window to be estimated and does not rely on a difference-based measure that could amplify noise (Hedge et al., 2018). To use this alternative approach to test for an effect of typicality, we fit a linear mixed-effects model predicting proportion target looking during the critical window from typicality condition

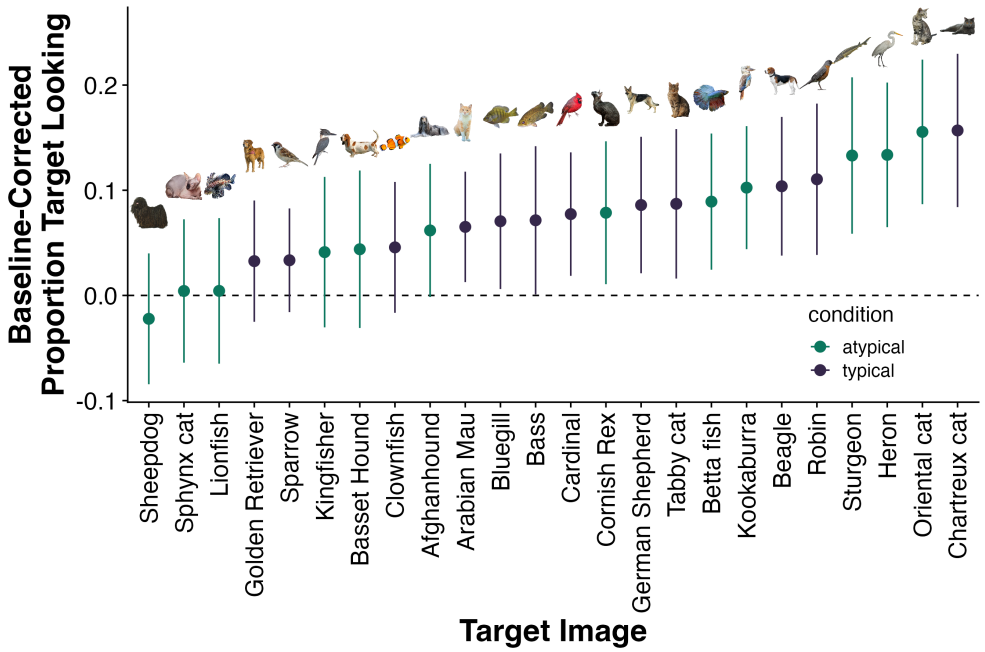

FIGURE 4 Average baseline-corrected proportion target looking for each target image, with color indicating typicality condition. Error bars represent 95% CIs.

(centered) while controlling for proportion target looking during the baseline window. We included by-participant and by-word random intercepts. As in the main analyses, we found no significant effect of typicality, $b = 0.01$, 95% CI [−0.01, 0.03], $t(3010) = 1.29$, $p = .20$. We also found no significant interaction between age and typicality using this modeling approach ($p = .49$).

Exploring the influence of both target and distractor typicality on word recognition

Using the same modeling approach, we also investigated the role of both target typicality (treated as a continuous predictor based on adult ratings) and distractor typicality on word recognition. We fit a linear mixed-effects model predicting proportion target looking during the critical window from target typicality ratings (z-scored), distractor typicality ratings (z-scored), and their interaction, while also controlling for proportion target looking during the baseline window. The model included by-participant and by-item random effects, as well as a by-participant random slope for the interaction between target and distractor typicality (more complex random-effects structures yielded a singular fit, albeit with comparable results). We found a significant interaction between target and distractor typicality, $b = -0.03$, 95% CI [−0.06, −0.01], $t(64.42) = -3.24$, $p = .002$. More typical distractors were associated with higher proportion target looking during the critical window in general ($b = 0.04$, 95% CI [0.02, 0.06], $t(2908) = 3.64$, $p < .001$), and this effect was stronger when the target was atypical (see Figure S7 in the Supporting Information for a visualization of this effect). We also found a similar interaction between target typicality and distractor typicality in predicting baseline-corrected proportion target looking in a linear mixed-effects model (i.e., in a model specification consistent with the main analytic approach), $b = -0.03$, 95% CI [−0.06, −0.01], $t(66.59) = -2.52$, $p = .01$, although here there was no significant overall effect of distractor typicality on baseline-corrected target looking ($p = .86$).

DISCUSSION

Across age (14 to 18 months) and exemplar typicality, word recognition was remarkably robust. When prompted with familiar animal words, infants successfully looked to the target not only for typical exemplars of the animal category, but also for atypical exemplars. These results show that infants' word representations are relatively broad early in lexical development: 14- to 18-month-old infants successfully linked familiar words (e.g., *bird*) to exemplars (e.g., a kookaburra) judged to be highly atypical based on both adult norms and parental report of their infants' experiences. We found no evidence for differences in word recognition for typical versus atypical exemplars in our preregistered analyses, and while recognition accuracy generally increased with age for both typical and atypical exemplars, we found no evidence for an interaction between age and typicality. At the same time, we were not able to draw strong conclusions about the absence (or presence) of a typicality effect for word recognition: Our equivalence test of differences in typicality was not significant, meaning that, despite a large infant sample, our data cannot rule out a typicality effect at least as large as our smallest effect size of interest ($d = 0.25$). Finally, we found no evidence

that individual differences in experience, as measured by parent-reported individual-specific typicality ratings, predicted word recognition accuracy. Overall, our results show that across age and individual differences in experience, infants develop the ability to recognize words represented by a broad range of referents.

Our results are largely consistent with a broad-to-narrow account of early word meanings. Specifically, infants readily extended animal nouns to include both typical and atypical exemplars and, in fact, showed no advantage for recognizing typical exemplars. It is worth highlighting the surprising nature of these results: our findings suggest that infants between 14 and 18 months link the word *bird* with a highly unusual exemplar such as a kookaburra—a bird species they are likely to never have encountered—approximately equally as well as they do with a highly typical exemplar (such as a robin). Infants' robust extension abilities, while surprising, suggest that word meanings are initially quite broad. These results contribute to mounting evidence that infants are able to connect familiar words with a wide range of exemplars even early on in development. For example, infants as young as 6 months of age have demonstrated evidence of correctly extending body parts and toys to both familiar (e.g., their own hand) and unfamiliar referents (e.g., a stranger's hand; Campbell & Hall, 2022; Garrison et al., 2020). Thus, converging evidence from multiple infant studies support a broad-to-narrow trajectory in the development of word meanings across several age ranges and category domains. Moreover, while several previous studies have reported significant typicality effects on word recognition (Meints et al., 1999; Perry & Saffran, 2017; Southgate & Meints, 2000), a bird's-eye view of the effect sizes documented in the literature suggests a consistent overarching pattern: typicality likely has, at most, a small effect on word recognition (see Supporting Information S6; Tables S3 and S3B for an overview). For instance, in a recent, well-powered study (N=287), 17- to 42-month-olds were equally accurate in identifying typical and atypical exemplars in a forced choice pointing task with three alternatives (Kucker et al., 2023), and even studies reporting significant typicality effects often find small effect size magnitudes (e.g., d ~0.2 in Perry & Saffran, 2017). We similarly found no evidence that infants' recognition of typical and atypical members differed in a series of robustness analyses varying analytic decisions and metrics of word comprehension (e.g., accuracy vs. reaction time). Together, these findings support an initially broad account of early word meaning by which infants extend labels to both typical and atypical exemplars.

While word recognition did not differ depending on the typicality of the target image, in exploratory analyses, we found evidence that the typicality of the distractor may exert influence on infants' looking to the named target. Specifically, higher distractor typicality was associated with higher proportion looking to the target during the critical window. This effect interacted with target typicality, such that the effect of distractor typicality was strongest in the atypical condition (when both target and distractor images were atypical). There are multiple possible explanations for these effects of distractor typicality. One possibility is that highly atypical distractors are visually salient and hold infants' attention, leading to a decrease in looking to the target specifically when the most atypical distractors are presented. This explanation is consistent with past work documenting a strong effect of visual salience on word recognition (e.g., Pomper & Saffran, 2019). Another possibility is that in the atypical condition, when both targets and distractors are atypical in general, it becomes especially important for infants to be able to reject the distractor as a possible referent of the target word. This process of ruling out the distractor may be more difficult for highly atypical distractors and thus require additional processing (cf. Kucker et al., 2023; Meints et al., 1999). While we cannot disentangle these alternative explanations given the current data, the results highlight the important role that both targets and distractors play in shaping infants' visual attention during word recognition.

Our study provides a snapshot of infant word meaning extension between 14 and 18 months of age. This choice of age range leaves open several possibilities concerning the nature of word meanings at both younger and older ages, and therefore different possible trajectories in the development of word meaning expertise. For one, our data cannot fully rule out the hypothesis that word meanings may in fact begin narrow at younger ages: perhaps infants younger than 14 months show a benefit in recognizing typical exemplars that disappears over the course of the first half of their second year (see also Meints et al., 1999). However, several features of our design and findings decrease the likelihood of this interpretation. We chose our age range to match the earliest ages at which we expected to robustly measure word recognition for our target labels (at least for typical exemplars), grounded in our pilot data (see Supplementary Material S7). Moreover, there was no age by typicality interaction in our current data, and related work suggests small—if any—effects of object familiarity even at 12 months of age (Garrison et al., 2020). A related question is whether and when typicality effects emerge at later ages in infants' word recognition. The fact that exemplar typicality exerts a strong influence on adults' (Rosch, 1978) and children's (Jerger & Damian, 2005) processing of familiar categories suggests that the influence of typicality should increase with age, which in turn would be consistent with a broad-to-narrow trajectory in the development of word meanings. Extending the current design to older ages could help clarify when and how knowledge of underlying category structure comes to structure infants' lexical processing.

Infants' early word meaning expertise gives rise to a key question: What early experiences facilitate broad

word extension? In the present sample, infants recognized both typical and atypical exemplars with seemingly minimal influence from their own experiences with the specific category members, as reported by their caregivers. Contrary to our a-priori prediction, this result suggests that a child with a pet sphynx cat, for example, was not more accurate at recognizing a sphynx as a referent for *cat* than a tabby cat. However, the caregiver survey employed in the current study is limited in several ways. For one, caregivers are likely to be imperfect reporters of their own child's familiarity with specific referent images, and the validity of caregiver judgments for individual infants' experience is unclear. For another, the survey was a coarse measure of early experience focused on a general judgment of the "typicality" of experience that does not capture the full breadth and variability of individual infants' experience with these familiar animal categories.

The literature, however, points to variability as a critical component of infants' early category and language development. Specifically, variability in the labeled exemplars that infants experience promote retention (Twomey et al., 2014) and generalization (Perry et al., 2010). These variable experiences likely highlight diagnostic category features (Althaus & Plunkett, 2016), allowing for robust early word extension to both familiar exemplars and novel, atypical exemplars. Self-generated object variability, in particular, may be important for extracting invariant category features (Sloane et al., 2019). Yet, despite growing evidence that early experiences facilitate infants' robust extension, the nature of early experiences remains largely unspecified and understudied. Recent work documenting the richness of day-to-day experience with familiar objects and labels (e.g., Clerkin & Smith, 2022) and methods for collecting dense datasets of early visual experience (Sullivan et al., 2021) pave a path forward for documenting infants' experience with common categories and determining what features of these experiences allow infants to generalize word meanings broadly from a young age.

A second open question is the degree to which robust word recognition for a wide range of category exemplars reflects lexically specific processing, infants' underlying (non-linguistic) category knowledge, or both. In the current task, we cannot distinguish lexical processing from visual identification and processing of category exemplars, because visual and linguistic processing were integrated in the task design. Evidence from the visual category development literature suggests that infants in our current age range are highly sensitive to the similarity and typicality of exemplar members, categorizing prototypical exemplars of a category more easily than atypical exemplars by 13 months (Bauer et al., 1995). Infants are also sensitive to the distribution and frequency of exposure to specific exemplars even before the end of their first year of life (Kovack-Lesh et al., 2014; Oakes & Spalding, 1997). One explanation of our findings consistent with past literature is that infants are sensitive to differences in typicality between exemplars, but that these differences are not sufficient to incur a substantial cost in linking a familiar word with a (even highly atypical) referent. In the looking-while-listening task used in the current study, infants only needed to recognize typical or atypical category members as better candidate target referents compared to distractor exemplars drawn from an alternative category to succeed. Thus, a potential conclusion is that infants' representations of both typical and atypical exemplars may be good enough to support similar lexical processing in constrained visual contexts, despite emerging typicality structure in infants' visual categories. One way to test this interpretation would be to combine measurements of lexical processing with non-linguistic tasks tapping into infants' perceptual categorization of category exemplars (e.g., Quinn & Tanaka, 2007; Robinson, 2002). Multi-measure approaches could provide a richer window into underlying perceptual and linguistic representations supporting broad generalizations of word meanings.

Using Lookit to study infant looking time and word knowledge

Our results demonstrate the feasibility and promise of collecting infant looking data in unmoderated online sessions using platforms such as Lookit (Scott & Schulz, 2017). This study adds to the emerging literature demonstrating the possibility of measuring infant looking time in both moderated (Chuey et al., 2021) and unmoderated (Lo et al., 2023) online test sessions. Our results demonstrate that word recognition, including information about the timecourse of looking, can be measured robustly in unmoderated sessions, even at relatively young ages. Unmoderated online testing has several potential benefits for future data collection efforts. It is potentially more convenient and requires less time for many participating families, allowing researchers to collect much larger samples in a shorter period than is typical in infant research. Unmoderated testing sessions also could make it easier to collect more data points from individual infants by spreading trials across multiple short testing sessions. In the current study, we were able to double the number of trials collected by testing infants across multiple sessions. Data retention was high from session 1 to session 2 (78%), providing encouraging initial returns on this approach's feasibility. The possibility to collect more trials per infant is especially promising for studying individual differences, allowing researchers to collect denser samples and, therefore, derive more reliable individual-level estimates (Byers-Heinlein et al., 2022).

Despite the immense promise of unmoderated online data collection, there remain several key challenges when conducting research online that do not

arise in the lab. First, there is a high degree of variability in video quality, due to variation in users' technical set-up (e.g., users' webcams) and in internet quality. For example, average video frame rates varied widely across families (~6–60 Hz) and within individual sessions. Variability in frame rates necessarily introduces variation in the timing precision, posing challenges for timing-sensitive analyses such as reaction time estimates (see also Bacon et al., 2021). Relatedly, it is difficult to retain precise control over timing information in online settings, for a variety of reasons, including limitations of the Lookit platform, differences across users' hardware and software that delivers stimuli, and variability in internet connection quality. Therefore, collecting infant looking time data using Lookit may be better suited to certain kinds of research questions and analytic approaches. For example, in our current design, the central manipulation (typicality) was within-subjects, and the main analyses focused on overall looking time across a long time window, features that both mitigate concerns about variability in timing and timing precision across participants. A second, related consequence of variability in video quality and testing contexts is that offline coding proved to be substantially more difficult—and time-consuming—than offline coding of in-laboratory data. One promising development that mitigates this concern is the recent rapid advancement in automated gaze-coding, in particular the iCatcher+ system (Erel et al., 2023).

Limitations

There are several important limitations to the current work. First, the design was optimized for detecting whether infants are able to recognize atypical exemplars, not necessarily for highlighting differences between typical versus atypical exemplars. Alternative designs could potentially reveal latent differences in how infants link words with typical versus atypical exemplars. For example, one approach could pit typical versus atypical exemplars of the same category against one another on critical trials, creating competition between more or less typical exemplars during word recognition (for a similar approach see Southgate & Meints, 2000). More generally, exploring differences in how tasks establish competition between typical and atypical exemplars may be a fruitful direction for understanding variability in typicality effects across word recognition studies. Second, our results are constrained by our use of a small set of familiar words limited to one semantic domain (i.e., animals). We chose these words and this semantic domain in order to ensure that labels were familiar and engaging for infants in our age range, and to control for potential saliency differences, since infants tend to attend to animate items more than inanimate items. Moreover, using a small item set allowed us to collect a larger number of trials per item and infant, in order to test individual differences. However, this small set of words also limits our ability to generalize these results to other semantic domains; a broader range of labels could help provide a more representative picture of typicality effects in infants' overall vocabulary. Finally, while the sample we collected through Lookit varied more widely across several demographic characteristics (e.g., race/ethnicity) compared to our typical in-lab sample, our sample was largely limited to American English-speaking families with high educational attainment ($n = 80$ had a bachelor's degree or higher), which may limit generalizability of these results to other languages and cultural contexts. In particular, highly educated families may have greater access to experiences with animals (e.g., zoo and aquariums) that may support broad extension in this particular set of words. While a work in progress, we view efforts to expand infant testing online and beyond the lab as offering a path to improving our ability to recruit more representative samples (Bacon et al., 2021).

CONCLUSION AND FUTURE DIRECTIONS

Infants robustly recognize words across a wide range of exemplars by 14 months, and word recognition remains comparable for typical and atypical exemplars at least until 18 months. These results are more consistent with a broad-to-narrow account of the development of early word meanings for infants in our age range and suggest that infants extend novel word meanings broadly at an age when word recognition is beginning to increase in robustness (Bergelson, 2020). Our findings point to fruitful directions for future research that explore the lexical-semantic representations and early category experiences that support forming broad generalizations and how infants continue to refine their understanding of word meanings across development. The current results also chart a course for future work taking an individual differences approach to the development of word meanings (Wojcik et al., 2022). Studying when and how variation in experience shapes variation in word meanings could yield fruitful insights into sources of changes in infants' lexical expertise, especially given recent evidence revealing a surprising amount of individual variation in how adults represent word meanings (Wang & Bi, 2021). Being a word meaning expert as an adult means having both highly refined and—at least to some extent—idiosyncratic meaning representations. While our parent-report measure of individual experience was not predictive of word recognition in the current study, the parent-report data also indicated wide-ranging individual differences in infants' experience with familiar categories. New methods for collecting dense measurements of infants'

day-to-day experiences offer paths for developing rich, valid measures of variability in category experience (Clerkin & Smith, 2022; De Barbaro & Fausey, 2022; Sullivan et al., 2021). Studying how and when individual differences in experience shape infants' word meanings offers a promising path forward for understanding what drives growth and change in our developing understanding of word meaning.

ACKNOWLEDGMENTS

This work was supported by NSF-GRFP DGE-1747503 to MZ and grants from the NICHD awarded to MZ (F32HD110174), JS (R37HD037466), and the Waisman Center (U54 HD090256). We would like to thank the anonymous reviewers who provided valuable comments on both the Stage 1 and Stage 2 manuscript that helped greatly improve the paper.

DATA AVAILABILITY STATEMENT

This is a Stage 2 Registered Report. The Stage 1 manuscript was accepted in principle at Child Development. The approved Stage 1 protocol was preregistered at https://osf.io/rqbtn. The stimuli, experimental materials, data, and analytic code that support this study are openly available in a publicly accessible repository on the OSF at https://osf.io/3t8gf/.

ORCID

Haley Weaver https://orcid.org/0000-0002-6849-0785
Martin Zettersten https://orcid.org/0000-0002-0444-7059

REFERENCES

Althaus, N., & Plunkett, K. (2016). Categorization in infancy: Labeling induces a persisting focus on commonalities. *Developmental Science*, *19*, 770–780. https://doi.org/10.1111/desc.12358

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5–6), 509–559. https://doi.org/10.1177/0142723719869731

Arias-Trejo, N., & Plunkett, K. (2010). The effects of perceptual similarity and category membership on early word-referent identification. *Journal of Experimental Child Psychology*, *105*(1–2), 63–80. https://doi.org/10.1016/j.jecp.2009.10.002

Bacon, D., Weaver, H., & Saffran, J. (2021). A framework for online experimenter-moderated looking-time studies assessing infants' linguistic knowledge. *Frontiers in Psychology*, *12*, 703839. https://doi.org/10.3389/fpsyg.2021.703839

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrett, M. D. (1986). Early semantic representations and early word-usage. In S. A. Kuczaj & M. D. Barrett (Eds.), *The development of word meaning* (pp. 39–67). Springer.

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

Bauer, P. J., Dow, G. A., & Hertsgaard, L. A. (1995). Effects of prototypicality on categorization in 1- to 2-year-olds: Getting down to basic. *Cognitive Development*, *10*, 43–68. https://doi.org/10.1016/0885-2014(95)90018-7

Beckage, N., Smith, L. B., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS One*, *6*, e19348. https://doi.org/10.1371/journal.pone.0019348

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*, 142–149. https://doi.org/10.1111/cdep.12373

Bergelson, E., & Aslin, R. (2017). Semantic specificity in one-year-olds' word comprehension. *Language Learning and Development*, *13*, 481–501. https://doi.org/10.1080/15475441.2017.1324308

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*, 1996–2009. https://doi.org/10.1111/cdev.13079

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied PsychoLinguistics*, *37*, 1461–1476. https://doi.org/10.1017/S0142716416000060

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer*. http://www.praat.org

Bornstein, M. H., & Arterberry, M. E. (2010). The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute, and group versus individual analyses. *Developmental Psychology*, *46*, 350–365. https://doi.org/10.1037/a0018411

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Lexical leverage: Category knowledge boosts real-time novel word recognition in 2-year-olds. *Developmental Science*, *19*(6), 918–932. https://doi.org/10.1111/desc.12343

Borovsky, A., & Peters, R. E. (2019). Vocabulary size and structure affects realtime lexical recognition in 18-month-olds. *PLoS One*, *14*, 1–21. https://doi.org/10.1371/journal.pone.0219290

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*, 389–411. https://doi.org/10.1037/met0000159

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS One*, *5*, e10773. https://doi.org/10.1371/journal.pone.0010773

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*, e2296. https://doi.org/10.1002/icd.2296

Campbell, J., & Hall, D. G. (2022). The scope of infants' early object word extensions. *Cognition*, *228*, 105210. https://doi.org/10.1016/j.cognition.2022.105210

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology*, *12*, 734398. https://doi.org/10.3389/fpsyg.2021.734398

Clerkin, E. M., & Smith, L. B. (2022). Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of Sciences of the United States of America*, *119*, e2123239119. https://doi.org/10.1073/pnas.2123239119

Colunga, E., & Sims, C. E. (2017). Not only size matters: Early-talker and late-talker vocabularies support different word-learning biases in babies and networks. *Cognitive Science*, *41*, 73–95. https://doi.org/10.1111/cogs.12409

De Barbaro, K., & Fausey, C. (2022). Ten lessons about infants' everyday experiences. *Current Directions in Psychological Science*, *31*, 28–33. https://doi.org/10.1177/0963721421105953

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987–1006. https://doi.org/10.3758/s13428-018-1115-7

Dink, J. W., & Ferguson, B. (2015). *eyetrackingR*. http://www.eyetracking-r.com

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, *65*, 903–917. https://doi.org/10.2307/1131427

Emberson, L. L., & Rubinstein, D. Y. (2016). Statistical learning is constrained to less abstract patterns in complex sensory input (but not the least). *Cognition*, *153*, 63–78. https://doi.org/10.1016/j.cognition.2016.04.010

Erel, Y., Shannon, K. A., Chu, J., Scott, K., Kline Struhl, M., Cao, P., Tan, X., Hart, P., Raz, G., Piccolo, S., Mei, C., Potter, C., Jaffe-Dax, S., Lew-Williams, C., Tenenbaum, J., Fairchild, K., Bermano, A., & Liu, S. (2023). iCatcher+: Robust and automated annotation of infants' and young children's gaze behavior from videos collected in laboratory, field, and online studies. *Advances in Methods and Practices in Psychological Science*, *6(2)*, 1–23. https://doi.org/10.1177/25152459221147250

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*, 95–115. https://doi.org/10.1017/S0142716400001053

Ferguson, B., & Waxman, S. (2017). Linking language & categorization in infancy. *Journal of Child Language*, *44*, 527–552. https://doi.org/10.1017/S0305000916000568

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In A. Fernald, R. Zangl, A. L. Portillo, & V. A. Marchman (Eds.), *Developmental psycholinguistics* (pp. 97–135). John Benjamins. https://doi.org/10.1075/lald.44.06fer

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press. https://doi.org/10.7551/mitpress/11577.001.0001

Garrison, H., Baudet, G., Breitfeld, E., Aberman, A., & Bergelson, E. (2020). Familiarity plays a small role in noun comprehension at 12–18 months. *Infancy*, *25*, 458–477. https://doi.org/10.1111/infa.12333

Goldberg, A. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Gruendel, J. M. (1977). Referential extension in early language development. *Child Development*, *48*, 1567–1576. https://doi.org/10.2307/1128520

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009). Longitudinal analysis of early semantic networks. *Psychological Science*, *20*, 729–739. https://doi.org/10.1111/j.1467-9280.2009.02365.x

Hirsh-Pasek, K., Golinkoff, R. M., Hennon, E. A., & McGuire, M. J. (2004). Hybrid theories at the frontier of developmental psychology: The emergentist coalition model of word learning as a case in point. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 173–204). MIT Press.

Hurley, K., & Oakes, L. M. (2018). Infants' daily experience with pets and their scanning of animal faces. *Frontiers in Veterinary Science*, *5*, 1–11. https://doi.org/10.3389/fvets.2018.00152

Huttenlocher, J., & Smiley, P. (1987). Early word meanings: The case of object names. *Cognitive Psychology*, *19*, 63–89. https://doi.org/10.1016/0010-0285(87)90004-1

Jerger, S., & Damian, M. F. (2005). What's in a name? Typicality and relatedness effects in children. *Journal of Experimental Child Psychology*, *92*, 46–75. https://doi.org/10.1016/j.jecp.2005.04.001

Kartushina, N., & Mayor, J. (2019). Word knowledge in six- to nine-month-old Norwegian infants? Not without additional frequency cues. *Royal Society Open Science*, *6*, 180711. https://doi.org/10.1098/rsos.180711

Kovack-Lesh, K. A., McMurray, B., & Oakes, L. M. (2014). Four-month-old infants' visual investigation of cats and dogs: Relations with pet experience and attentional strategy. *Developmental Psychology*, *50*, 402–413. https://doi.org/10.1037/a0033195

Kucker, S., Braun, B. E., & Markham-Anderson, J. F. (2023). Margarita glasses and high heels: How attention to shape, age, and vocabulary impacts children's recognition of typical and atypical exemplars. *Child Development*, *94*, 603–616. https://doi.org/10.1111/cdev.13883

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*, 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*, 259–269. https://doi.org/10.1177/2515245918770963

Lo, C. H., Hermes, J., Kartushina, N., Mayor, J., & Mani, N. (2023). e-Babylab: An open-source browser-based tool for unmoderated online developmental studies. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02200-7

Mani, N., Johnson, E., McQueen, J. M., & Huettig, F. (2013). How yellow is your banana? Toddlers' language-mediated visual search in referent-present tasks. *Developmental Psychology*, *49*, 1036–1044. https://doi.org/10.1037/a0029382

Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H. M. (2018). Speed of language comprehension at 18 months old predicts school-relevant outcomes at 54 months old in children born preterm. *Journal of Developmental and Behavioral Pediatrics*, *39*, 246–253. https://doi.org/10.1097/DBP.0000000000000541

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*, 1–31. https://doi.org/10.1037/a0018130

Meints, K., Plunkett, K., & Harris, P. L. (1999). When does an ostrich become a bird? The role of typicality in early word comprehension. *Developmental Psychology*, *35*, 1072–1078. https://doi.org/10.1037/0012-1649.35.4.1072

Oakes, L. M., Coppage, D. J., & Dingel, A. (1997). By land or by sea: The role of perceptual similarity in infants' categorization of animals. *Developmental Psychology*, *33*, 396–407. https://doi.org/10.1037/0012-1649.33.3.396

Oakes, L. M., & Spalding, T. L. (1997). The role of exemplar distribution in infants' differentiation of categories. *Infant Behavior and Development*, *20*, 457–475. https://doi.org/10.1016/S0163-6383(97)90036-9

Perry, L. K., & Saffran, J. R. (2017). Is a pink cow still a cow? Individual differences in toddlers' vocabulary knowledge and

lexical representations. *Cognitive Science*, *41*, 1090–1105. https://doi.org/10.1111/cogs.12370
Perry, L. K., & Samuelson, L. K. (2011). The shape of the vocabulary predicts the shape of the bias. *Frontiers in Psychology*, *2*, 1–12. https://doi.org/10.3389/fpsyg.2011.00345
Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science*, *21*, 1894–1902. https://doi.org/10.1177/0956797610389189
Peters, R., & Borovsky, A. (2019). Modeling early lexico-semantic network development: Perceptual features matter most. *Journal of Experimental Psychology: General*, *148*, 763–782. https://doi.org/10.1037/xge0000596
Pomper, R., & Saffran, J. R. (2019). Familiar object salience affects novel word learning. *Child Development*, *90*(2), e246–e262. https://doi.org/10.1111/cdev.13053
Poulin-Dubois, D., & Sissons, M. E. (2002). Is this still called a dog? 18-month-olds' generalization of familiar labels to unusual objects. *Infant and Child Development*, *11*, 57–67. https://doi.org/10.1002/icd.256
Quinn, P. C., & Tanaka, J. W. (2007). Early development of perceptual expertise: Within-basic-level categorization experience facilitates the formation of subordinate-level category representations in 6- to 7-month-old infants. *Memory and Cognition*, *35*, 1422–1431. https://doi.org/10.3758/BF03193612
R Development Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/
Rescorla, L. A. (1980). Overextension in early language development. *Journal of Child Language*, *7*(2), 321–335. https://doi.org/10.1017/S0305000900002658
Robinson, C. W. (2002). *Comprehension of labels denoting typical and atypical exemplars: What is guiding lexical extensions?* The University of Toledo.
Rosch, E. (1978). Principles of categorization. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings* (pp. 189–206). MIT Press.
Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*, 4–14. https://doi.org/10.1162/opmi_a_00002
Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. https://doi.org/10.1177/0956797614567341
Sloane, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental Science*, *22*, e12816. https://doi.org/10.1111/desc.12816
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. https://doi.org/10.1111/1467-9280.00403
Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a Lexicon* (pp. 257–294). MIT Press.
Southgate, V., & Meints, K. (2000). Typicality, naming, and category membership in young children. *Cognitive Linguistics*, *11*, 5–16. https://doi.org/10.1515/cogl.2001.011
Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78. https://doi.org/10.1207/s15516709cog2901_3
Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, *5*, 20–29. https://doi.org/10.1162/opmi_a_00039
Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*, 480–484. https://doi.org/10.1111/1467-9280.00485
The jamovi project. (2020). *jamovi (1.2)*. https://www.jamovi.org
Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*, 172–175. https://doi.org/10.1111/1467-9280.00127
Twomey, K. E., Lush, L., Pearce, R., & Horst, J. S. (2014). Visual variability affects early verb learning. *British Journal of Developmental Psychology*, *32*, 359–366. https://doi.org/10.1111/bjdp.12042
Wang, X., & Bi, Y. (2021). Idiosyncratic tower of babel: Individual differences in word-meaning representation increases as word abstractness increases. *Psychological Science*, *31*, 1617–1635. https://doi.org/10.1177/09567976211003877
Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*, 258–263. https://doi.org/10.1016/j.tics.2009.03.006
Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology*, *29*, 257–302. https://doi.org/10.1006/cogp.1995.1016
Willits, J. A., Wojcik, E. H., Seidenberg, M. S., & Saffran, J. R. (2013). Toddlers activate lexical semantic knowledge in the absence of visual referents: Evidence from auditory priming. *Infancy*, *18*, 1–23. https://doi.org/10.1111/infa.12026
Wojcik, E., Zettersten, M., & Benitez, V. (2022). The map trap: Why and how word learning research should move beyond mapping. *WIREs Cognitive Science*, *13*, e1596. https://doi.org/10.1002/wcs.1596
Wojcik, E. H. (2018). The development of lexical–semantic networks in infants and toddlers. *Child Development Perspectives*, *12*, 34–38. https://doi.org/10.1111/cdep.12252
Wojcik, E. H., & Saffran, J. R. (2013). The ontogeny of lexical networks: Toddlers encode the relationships among referents when learning novel words. *Psychological Science*, *24*, 1898–1905. https://doi.org/10.1177/0956797613478198
Yin, J., & Csibra, G. (2015). Concept-based word learning in human infants. *Psychological Science*, *26*, 1316–1324. https://doi.org/10.1177/0956797615588753
Zettersten, M., Bergey, C. A., Bhatt, N. S., Boyce, V., Braginsky, M., Carstensen, A., deMayo, B., Kachergis, G., Lewis, M., Long, B., MacDonald, K., Mankewitz, J., Meylan, S., Saleh, A. N., Schneider, R. M., & Tsui, A. S. M. (2021). Peekbank: Exploring children's word recognition through an open, large-scale repository for developmental eye-tracking data. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Weaver, H., Zettersten, M., & Saffran, J. R. (2024). Becoming word meaning experts: Infants' processing of familiar words in the context of typical and atypical exemplars. *Child Development*, *00*, 1–21. https://doi.org/10.1111/cdev.14120