Word learning in 14-month-old monolinguals and bilinguals: Challenges and methodological opportunities

Ana Maria Gonzalez-Barrero[1]
Rodrigo Dal Ben[2]
Hilary Killam
Krista Byers-Heinlein

Department of Psychology, Concordia University, Canada
Centre for Research on Brain, Language and Music, Canada

Abstract: Infants can learn words in their daily interactions early in life, and many studies have demonstrated that they can also learn words from brief in-lab exposures. While most studies have included monolingual infants, less is known about bilingual infants' word learning and the role that language familiarity plays in this ability. In this study we examined word learning in a large sample (up to $N = 148$) of bilingual and monolingual 14-month-olds using a preferential looking paradigm. Two novel words were presented within sentence frames in one language (single-language condition) or two languages (dual-language condition). We predicted that infants would learn both words, and would exhibit better learning when they were more familiar with the sentence frame language. Using a traditional analytic approach (*t*-tests) and a standard linear regression, we found weak evidence that children learned one of the two words. However, contrary to our prediction, in a minority of conditions infants may have learned better when stimuli were presented in sentence frames in a less familiar language. We also conducted updated analyses using mixed-effects linear regression models, which did not support the conclusion that infants learned any of the words they encountered, regardless of the familiarity of the sentence frame language. We discuss these results in relation to prior work and suggest how open science practices can contribute to more reliable findings about early word learning.

Keywords: word learning; infants; bilingualism; open science.

Corresponding author: Ana Maria Gonzalez-Barrero, Department of Psychology, Concordia University, 7141 Sherbrooke St. West, Montreal, QC, H4B 1R6, Canada. E-mail: ana.gonzalez@mail.mcgill.ca

ORCID ID(s): Ana Maria Gonzalez-Barrero (https://orcid.org/0000-0002-2120-6329), Rodrigo Dal Ben (https://orcid.org/0000-0003-2185-8762), Hilary Killam (https://orcid.org/0000-0002-3080-0610), Krista Byers-Heinlein (https://orcid.org/0000-0002-7040-2510)

Citation: Gonzalez-Barrero, A. M., Dal Ben, R., Killam, H., & Byers-Heinlein, K. (2023). Word learning in 14-month-old monolinguals and bilinguals: Challenges and methodological opportunities. *Language Development Research, 3*(1), 277—317. http://doi.org/10.34842/3vw8-k253

[1] Ana Maria Gonzalez-Barrero is now at School of Communication Sciences and Disorders, Faculty of Health, Dalhousie University, Nova Scotia, Canada.
[2] Rodrigo Dal Ben is now at Ambrose University, Psychology Program, Alberta, Canada.

Introduction

Word learning is a complex process that begins to unfold over the first two years of life. Past research, using a variety of experimental designs, has provided important insights into word learning in both monolingual and bilingual infants (e.g., Fennell & Byers-Heinlein, 2014; Graf Estes, 2014; Kalashnikova et al., 2018; Mattock et al., 2010; Schafer & Plunkett, 1998; Singh et al., 2018; Taxitari et al., 2020; Werker et al., 1998; Woodward et al., 1994). However, a new understanding of research best practices highlights the limitations of our traditional methodological and statistical approaches (Bergmann et al., 2018; Oakes, 2017). In this manuscript, we present a case study using both traditional and more sensitive analytic techniques to examine word learning in the lab with a large sample of 14-month-old monolinguals and bilinguals ($N = 148$). Following prior research (Fennell & Byers-Heinlein, 2014; Fennell & Waxman, 2010), we presented infants with novel words embedded into sentence frames, and then tested their learning in a looking time paradigm. We discuss whether infants of this age were able to learn new words in this context, and how our traditional research practices in early language acquisition can be improved to produce more reliable and reproducible findings.

Word Learning

Researchers have used experimental tasks to study word learning for more than 40 years (e.g., Carey & Bartlett, 1978). Dozens of studies have shown that infants and children can learn new words in the lab (Dal Ben et al., 2019; Gonzalez-Gomez et al., 2013; Kalashnikova et al., 2015; Ramon-Casas et al., 2009; Shukla et al., 2011; Singh et al., 2018; Tsui et al., 2019; Tsuji et al., 2020b; Yu & Smith, 2011). They have also shown that infants' ability to make initial mappings between words and their referents, as well as their ability to retain these mappings, are affected by multiple factors related to the nature of the word learning task, including the particular stimuli used and how the words are encountered (Burnham et al., 2018; Hirsh-Pasek et al., 2000; Kucker et al., 2015; McMurray et al., 2012; Werker & Curtin, 2005), as well as other factors like language background (Tsui et al., 2019) and vocabulary size (Werker et al., 2002). Word learning skills also seem to improve with age (Frank et al., 2021). As early as 6 months old, everyday language experience supports infants' ability to associate labels with referents such as food and body parts (Bergelson & Swingley, 2012), and at this age there is also evidence that infants can learn new words in the lab (Shukla et al., 2011). Some have argued, however, that there is a qualitative change in infants' word comprehension abilities that occurs just after their first birthday (Bergelson, 2020). Indeed, there are many more reports of infants aged 12 months and older showing successful word learning in the lab than reports of younger children (e.g., Graf Estes et al., 2007; Lany, 2014; Yin & Csibra, 2015). Our study tested infants at an age where basic laboratory word learning is thought to be relatively robust: 14 months.

Successful word learning in the lab has been reported in different contexts. For example, Woodward et al. (1994) found that 13-month-old monolingual infants mapped a novel word to its referent after only nine encounters with the word–referent pair presented by a live experimenter. More recently, Chen and colleagues (2020) used a similar paradigm and found that monolingual 20-month-olds could learn a native and a foreign word after only 6 encounters with the word–referent pair. Lab studies using more stripped-down tasks (i.e., without live social interaction) have also shown successful word learning. For instance, 15-month-olds learned two novel words in a preferential looking paradigm where isolated novel words (e.g., "*bard*") and pictures of novel objects were paired, with no social agents or social support (Schafer & Plunkett, 1998). In addition, using the Switch task, wherein infants are habituated to two novel word–object pairings and then presented with a mismatch at test, similar results were found with 14-month-old monolinguals (Werker et al., 1998) and bilinguals (Byers-Heinlein et al., 2013).

For slightly older monolingual (15-month-olds, Fennell & Waxman, 2010) and bilingual (17-month-olds, Fennell & Byers-Heinlein, 2014) infants, sentence frames have been shown to enhance word learning and recognition. For instance, monolingual 18-month-olds showed faster recognition of familiar words presented in sentence frames than in isolation (Fernald & Hurtado, 2006). Moreover, while 14-month-olds often find minimal pair word learning challenging (Stager & Werker, 1997), monolingual 14-month-olds successfully mapped a minimal pair (*bin* and *din*) to objects during the Switch task when the words were embedded in sentence frames (Fennell & Waxman, 2010). Similarly, 16-month-old bilinguals, whose languages shared linguistic similarities (e.g., French and Spanish), mapped a minimal pair (*tola* and *dola)* to objects when words were embedded in sentence frames and presented in a live interaction experiment (Havy et al., 2016). Seventeen-month-old monolinguals and French–English bilinguals also learned minimal pair labels (*kem* and *gem*) embedded into sentences that were produced by a speaker that matched their language background (Fennell & Byers-Heinlein, 2014). Sentence frames may support word learning by providing familiar linguistic context, highlighting the referential nature of the word learning task, and decreasing infants' cognitive load. This information might be particularly useful for bilingual infants, who may use sentential frames to navigate between languages (Fennell & Byers-Heinlein, 2014; Fernald & Hurtado, 2006; Havy et al., 2016).

Current Study

The current study extended previous research by investigating word learning just after infants' first birthdays. We asked if 14-month-old infants would successfully learn new words embedded in sentence frames. Moreover, we were interested in the role of infants' language background, specifically whether they were growing up in a monolingual or a bilingual environment. Despite deploying similar mechanisms for word learning (Byers-Heinlein et al., 2013; Byers-Heinlein & Werker, 2009; Kandhadai et al., 2017), infants growing up bilingual are exposed to unique input that may impact their language development (Fennell et al., 2007; Graf Estes & Hay, 2015). For instance, bilingual infants often hear interlocutors alternate between two languages in the same contexts (Place & Hoff, 2011), especially when bilingual parents teach new words (Byers-Heinlein, 2013; Kremin et al., 2022). Recent studies indicate that some types of language alternation make word learning challenging (Byers-Heinlein et al., 2022), although other evidence suggests that many instances of parental code-switching are supportive for learning (Kremin et al., 2022).

We presented monolingual and bilingual infants with novel words embedded in sentence frames that differed in linguistic familiarity. Specifically, we presented 14-month-olds pictures of novel objects paired with the dissimilar-sounding novel words "*kem*" and "*bos*" embedded in English and/or French sentence frames. Our study had two training conditions. In the single-language condition, both words were presented in the same language (either in English or in French sentence frames), and in the dual-language condition each word was presented in a different language (one in English sentence frames and one in French sentence frames). After training, infants were tested in a preferential looking paradigm, where they saw both novel objects side-by-side and heard one of the words in isolation. Infants came from one of three backgrounds: (a) monolingual English *or* French, (b) bilingual English *and* French, or (c) bilingual English *or* French *and* another language. That is, all infants had exposure to one or both of the sentence frame languages (English and French), but to varying degrees, as bilingual infants are rarely perfectly balanced in their exposure to each of their languages (i.e., they typically have a dominant and a non-dominant language; Byers-Heinlein et al., 2019). By including infants from these diverse language backgrounds, we could examine the effects of bilingualism as well as infants' familiarity with the sentence frame languages.

Given the bulk of evidence from the published literature that 14-month-old monolinguals and bilinguals can successfully learn new words in the lab, we expected that at least under some conditions, infants would also be successful in our task. More specifically, we expected that the more familiar infants were with the language of the sentence frame, the better they would learn the novel words. For instance, we expected that word learning would be easier for bilingual infants when they heard the sentence frame in their dominant rather than in their non-dominant language, and infants should have the most difficulty learning a new word embedded in foreign language sentence frames (e.g., an English monolingual infant hearing a French sentence).

Building on previous research showing that vocabulary size (Werker et al., 2002) and attention (Kannass & Oakes, 2008; Yu & Smith, 2012) can influence infants' word learning, we also investigated these additional variables to provide a more complete account of our findings. Thus, we also explored whether infants would show better learning as a function of how many words they knew in the sentence frame language (e.g., how many words they knew in English), and whether attention during the training phase contributed to successful word learning.

This project began in 2012, and the combination of different language backgrounds and conditions was originally conceptualized as forming a set of 7 different experiments (see Table 1).[3] Following past studies, we had planned a sample size of 16 infants per condition (see Oakes, 2017, for evidence that this sample size is typical of many infant experiments, although a recent meta-analysis has revealed that this often yields underpowered experiments; Bergmann et al., 2018). However, after 7 years of data collection (2012–2019), and despite collecting data from 288 infants (many of whom ultimately had to be excluded from analyses, discussed further below), we were able to achieve our target sample for only some of the experiments, and thus chose to terminate data collection. We note that by this point the last author of this paper (the Principal Investigator) was the only researcher still in the lab from the time the experiment began. A subset of these data from monolingual infants—who were substantially easier to recruit—was published by da Estrela and Byers-Heinlein (2016), who designed the experimental approach and created the stimuli. They reported an experiment under which monolinguals learned the novel words, as well as two experiments in which they

[3] Our original intention was to investigate word learning in French–English bilinguals, building from a series of word learning studies from that time (Fennell & Byers-Heinlein, 2014; Fennell et al., 2007; Mattock et al., 2010). However, recruitment of French–English bilinguals was slow, and we were turning away many interested families with other language backgrounds. We thus expanded our research design to collect data from monolinguals as well as bilingual infants learning French/English and an additional language. The categorization of infants as monolingual or bilingual (rather than taking a continuous approach to language exposure) was consistent with the literature at the time. We prioritized testing infants in the dual language condition (the first condition we designed), and additionally tested monolinguals and French–English bilinguals in the single-language condition. As there are many more monolinguals in our community than any of the groups of bilinguals, these infants were tested in the greatest number of conditions. It was expected that these infants' different relationships with the sentence frame languages might provide some additional insight into factors – such as familiarity – that could influence early word learning, while allowing us to increase the number of infants tested and accommodate a wider range of families.

failed to learn. In retrospect, we note that all of the studies reported back then were underpowered, which could lead to spurious findings (Oakes, 2017). We expand on this point in the Discussion.

Table 1. ***Examples of Infants' Familiarity with Sentence Frame Languages***

Language Group	Experiment Number	Language Background	Infants' Most Familiar Sentence Frame Language (e.g., English; *Look! It's the Bos*!)	Infants' Least Familiar Sentence Frame Language (e.g., French; *Regarde! C'est le Kem!)*
		Dual-Language Condition		
Bilinguals	1	L1 English L2 French	Dominant	Non-Dominant
	2	L1 English L2 Other	Dominant	Foreign
	3	L1 Other L2 English	Non-Dominant	Foreign
Monolinguals[a]	4	L1 English	Dominant/Native	Foreign
		Single-Language Condition		
Bilinguals	5	L1 English L2 French	Dominant	NA
Monolinguals	6	L1 English	Dominant/Native	NA
Monolinguals[a]	7	L1 French	NA	Foreign

Note. In these examples English is the most familiar language and French is the least familiar language. The relationships are reversed when French is the most familiar language. L1 = Infants' dominant (or native in the case of monolinguals) language; L2 = Infants' non-dominant language.

[a]Monolingual infants included in a prior study.

The experiments presented here were conceptualized before new approaches (e.g., large-scale collaborations; ManyBabies Consortium, 2020) and articles calling for better research practices were widely disseminated in the field of developmental psychology (Bergmann et al., 2018; Bishop, 2020; Oakes, 2017; Schott et al., 2019), although such ideas were being discussed in some other fields (Button et al., 2013; Ioannidis, 2005; John et al., 2012; Simmons et al., 2011). However, for different reasons, most notably the slow pace of infant data collection (especially with bilingual infants), we found ourselves analyzing our data after improved research practices were becoming more common in infant research and the field of bilingualism. This laid bare some problematic characteristics of our original approach, which would likely have characterized many published studies in the field: it was not pre-registered; it had small sample sizes per experimental group; planned statistical analyses focused on small individual experiments rather than the dataset as a whole; there was potential for undisclosed flexibility in the analysis; and monolingualism and bilingualism were defined categorically in a way that ultimately excluded many participants who were tested (see Byers-Heinlein, 2015; Kremin & Byers-Heinlein, 2021; Luk, 2015; Luk & Bialystok, 2013, for a longer discussion of categorical versus continuous approaches to bilingualism).

Our conundrum raises an important question for studies with a long gap between study planning and data analysis: when should researchers stick with their original plan that is consistent with the rest of the literature, and when should they use updated approaches such as combined analyses that yield larger sample sizes, advanced statistical methods, and open science practices? We have ultimately decided to take both paths at once, in order to better understand how we should conceptualize older versus newer research practices in the context of the literature on infant word learning. In what follows, we first present our planned analysis (which we refer to as the traditional approach) and then a re-analysis of our data using a more sensitive technique (which we refer to as the updated approach). Finally, we discuss how the use of traditional versus updated approaches can affect our conclusions about infant experimental word learning tasks, contributing to the discussion on how to improve practices in infant research.

Method

Analytic Approaches

We report two analytic approaches. In the traditional approach, we used one-sample, two-tailed *t*-tests against chance to test word learning for each experiment in each condition, following da Estrela and Byers-Heinlein (2016). In the updated approach, we analyzed all experiments in aggregate (the full sample), using mixed-effects models. Critically, both analytic approaches used the same window of analysis, which began

200ms after the onset of the first iteration of the target word and lasted until the end of the testing trial, 10000ms (see Design section). The 200-ms shift was to account for the time it takes infants to initiate an eye movement (Canfield et al., 1997). The total length of the analysis window was 6800ms. Both approaches were implemented in R (R Core Team, 2020) and all data and scripts are available at https://osf.io/upy7f.

Participants

A total of 288 infants were tested between August 2012 and July 2019. This study was conducted in Montreal, Canada, a multicultural city where a high proportion of children are raised in a bilingual environment (Schott et al., 2022). Following exclusion criteria from prior studies (Byers-Heinlein et al., 2021; Mattock et al., 2010; Tsuji et al., 2020b), we excluded infants born premature (i.e., < 37 weeks of gestation, $n = 10$), with low birth weight (i.e., < 2500 grams, $n = 11$), with major health issues ($n = 1$), and those who were too fussy or inattentive to complete the study (for example, children who cried extensively during the experiment were considered fussy and children who refused to look at the screen were considered inattentive; $n = 44$).

We also excluded infants due to technical problems (e.g., connection problems with the eye-tracker; $n = 17$), experimenter error ($n = 4$), parental interference during the experimental portion of the study ($n = 2$), and those without enough looking data obtained from testing trials ($n = 7$). We defined enough looking data as at least 750ms of looking time during the specified windows of analyses for testing trials (following da Estrela & Byers-Heinlein, 2016), to ensure at least minimal attention was paid to the task, thus we excluded trials with less than 750ms of total looking from our analyses.

In addition, we excluded bilingual infants who were not exposed to both languages from birth or for whom age of acquisition was not reported ($n = 29$), bilingual infants who did not meet the study's language criteria, only discovered once infants participated in the study and parents completed the detailed language exposure questionnaire (i.e., exposure to a second language did not reach at least 25%, $n = 26$; see Rocha-Hidalgo & Barr, 2022, for a discussion of bilingualism criteria used in infant studies), infants who were not exposed to the target languages ($n = 3$), or children who were regularly exposed to 3 languages ($n = 11$). Infants who did not have at least one testing trial with adequate looking data per target word (i.e., at least one valid testing trial for "*kem*" and one for "*bos*", $n = 13$) were excluded from the traditional approach. We return to the issue of this reduction of sample size due to exclusions in the Discussion.

The final sample for the traditional approach included 110 14-month-olds (age range: 13 months and 16 days – 15 months and 12 days, Mean: 14 months and 12 days, *SD*: 13.6

days; 57 females) from diverse language backgrounds. Monolingual infants (n = 50) were exposed to one language, either English or French, 90% of the time or more. Bilingual infants were exposed at least 25% of the time to each of two languages, and less than 20% to a third language. We included bilingual infants exposed to English *and* French (n = 35) and bilingual infants exposed to English *or* French and another language (n = 24). Bilingual infants in all studies were exposed to at least one of the sentence frame languages (English and French) since birth. Participants' demographic characteristics are presented in Table 2.

The final sample for the updated approach consisted of 148 infants[4], a 35% increase in sample size compared to the traditional approach. This included all infants from the traditional approach (n = 110). It also included infants who had been excluded from the traditional approach because their language exposure fell outside the criteria established for bilingualism or monolingualism, except for one infant who did not have at least 750ms of looking time during testing trials (n = 25). We also included the 13 infants who had been excluded from the traditional approach for not having at least one valid test trial per novel word. These additional infants could be included in the updated approach because we treated language exposure continuously rather than categorically, and because mixed effects models are able to handle missing data.

Stimuli

We used the same stimuli and general procedure as da Estrela and Byers-Heinlein (2016). All our stimuli are openly available at https://osf.io/g6nrv. The visual stimuli had been used in prior research examining word learning in monolingual and bilingual infants (Byers-Heinlein et al., 2013; Curtin et al., 2009; Fennell et al., 2007; Fennell & Waxman, 2010; Werker et al. 1998; Werker et al., 2002). The auditory stimuli were recorded in our lab, and were originally chosen from the stimuli used in other previous studies of minimal pair word learning in French–English bilinguals: *bos* had been used by Mattock et al. (2010) and *kem* had been used by Fennell and Byers-Heinlein (2014). The two words do not overlap in sound and contain phonemes that are produced similarly across Canadian French and Canadian English (for a more complete comparison of the

[4] Our final samples for both the traditional and updated approaches included 28 monolingual infants whose data were previously published as a subsample of this larger dataset (Experiment 1 and Experiment 2 from da Estrela and Byers-Heinlein, 2016), which we reanalyzed in Studies 4 and 7 (see Table 1). Total sample sizes from both the traditional (n = 110) and updated approach (N = 148) provide our study with more than 80% statistical power to detect moderate to low effect sizes like the one estimated by a meta-analysis of studies with 12- to 16-months-old infants learning words in the Switch Task (d = 0.33; Tsui et al., 2019). However, individually, experiments 1–7 (n = 10–19) were underpowered. The full power analysis is available at: https://osf.io/upy7f.

realization of the relevant speech sounds in each language, see Fennell & Byers-Heinlein, 2014; Mattock et al., 2010).

Table 2. ***Infant Demographic Characteristics by Language Group***

Language Group	*n*	Mean Age in Months *(SD)*	Age Range	Sex	*n* per Language Dominance	Mean % Language Exposure
Traditional Approach						
Mono-lingual	50	14m 13d *(13.3d)*	13m 17d–15m 7d	48% female	24 English native 26 French native	98% EN 98.5% FR
Bilingual English–French	36	14m 14d *(14.9d)*	13m 16d–15m 12d	56% female	20 English dominant 16 French dominant	65% EN & 33% FR 61% FR & 37% EN
Bilingual English or French & Other Language	24	14m 6d *(11.1d)*	13m 24d–14m 29d	58% female	5 English dominant 5 French dominant 14 dominant in another language (6 with English and 8 with French as non-dominant language)	64% EN & 36% OT 61% FR & 39% OT 64% OT & 36% EN or FR
Additional Infants – Updated Approach						
38 infants included in the updated approach		14m 10d *(12.9d)*	13m 22d–15m 12d	56% female	24 English dominant 14 French dominant 3 dominant in another language (1 with English and 2 with French as non-dominant language)	78% EN, 20% FR, & 2% OT 75% FR, 19% EN, & 6% OT 67% OT, 33% EN and/or FR

Note. *M* = mean; *SD* = Standard Deviation, m = months, d = days, EN = English, FR = French, OT = Other Language. The percentage of language exposure does not add to 100% in some cases, since some infants in monolingual or bilingual groups had a small amount of exposure to other languages.

Visual stimuli consisted of two novel objects: a crown shape and a molecule shape (Figure 1). Target words (*bos* and *kem*) were presented embedded in English and/or French sentence frames (training) or in isolation (test). Across experimental conditions the molecule shape was always labelled with the novel word *kem* and the crown shape was always labelled with the novel word *bos*. Three unique tokens/recordings of each target word were used during training, always favouring the natural flow of the auditory stimuli. Identical tokens for the target words were used across all conditions, on both English and French sentences, which was accomplished through cross-splicing tokens that were pronounced in a way that was neither distinctly English nor distinctly French (according to an informal survey of speakers of each language). There were 3 sentence frames used in English ("Look, it's the ___!"… "Do you see the ___?"… "I like the ___!") and 3 in French ("Regarde, c'est le ___!"… Vois-tu le ___?"… "J'aime le ___!"). The novel words (i.e., *bos* and *kem*) were always presented in a sentence-final position to increase their salience (Echols & Newport, 1992; Fernald & Hurtado, 2006), and to support infants in segmenting out the target word even when the sentence frame was less familiar (Seidl & Johnson, 2006). Sentences were matched on length and prosody to minimize differences across the stimuli and were selected to ensure that the stimuli sounded as natural as possible. There were no sentence frames used in the test phase, and so the exact same recordings were used for French and English. The tokens used for the test phase were different from the ones used in the training phase. All stimuli were recorded by a native bilingual English–French female using infant-directed speech.

Auditory and visual stimuli were combined into videos to create training and test trials. Training trials presented the target object looming against a black background. The visual stimulus appeared in silence for the first 1.5 seconds, followed by 8 seconds where it was accompanied by an auditory stimulus with the target novel word embedded in either a French or English sentence, followed finally by 1.5 seconds of silence. Three sentences were presented during each training trial (e.g., "Look, it's the *kem*!"… "Do you see the *kem*?"… "I like the *kem*!"), with an interval of 1.5 seconds of silence between them. The duration of each training trial was approximately 11 seconds. During Test trials, visual stimuli (i.e., the crown-shaped and molecule-shaped objects) were presented side by side for the entire duration of the trial (≈ 10 seconds). During the first 3 seconds, visual stimuli were presented in silence, then isolated target words were played three times (e.g., "*Kem*!"…"*Kem*!"… "*Kem*!") with 1.5 seconds of pause between repetitions. Visual stimuli remained on the screen for a final 1.5 seconds of silence, before a new test trial began. Test trials were presented in one of four counterbalanced orders, which were identical across conditions. Stimuli are available at https://osf.io/g6nrv. Figure 1 shows the stimuli and timeline during an example training and an example test trial.

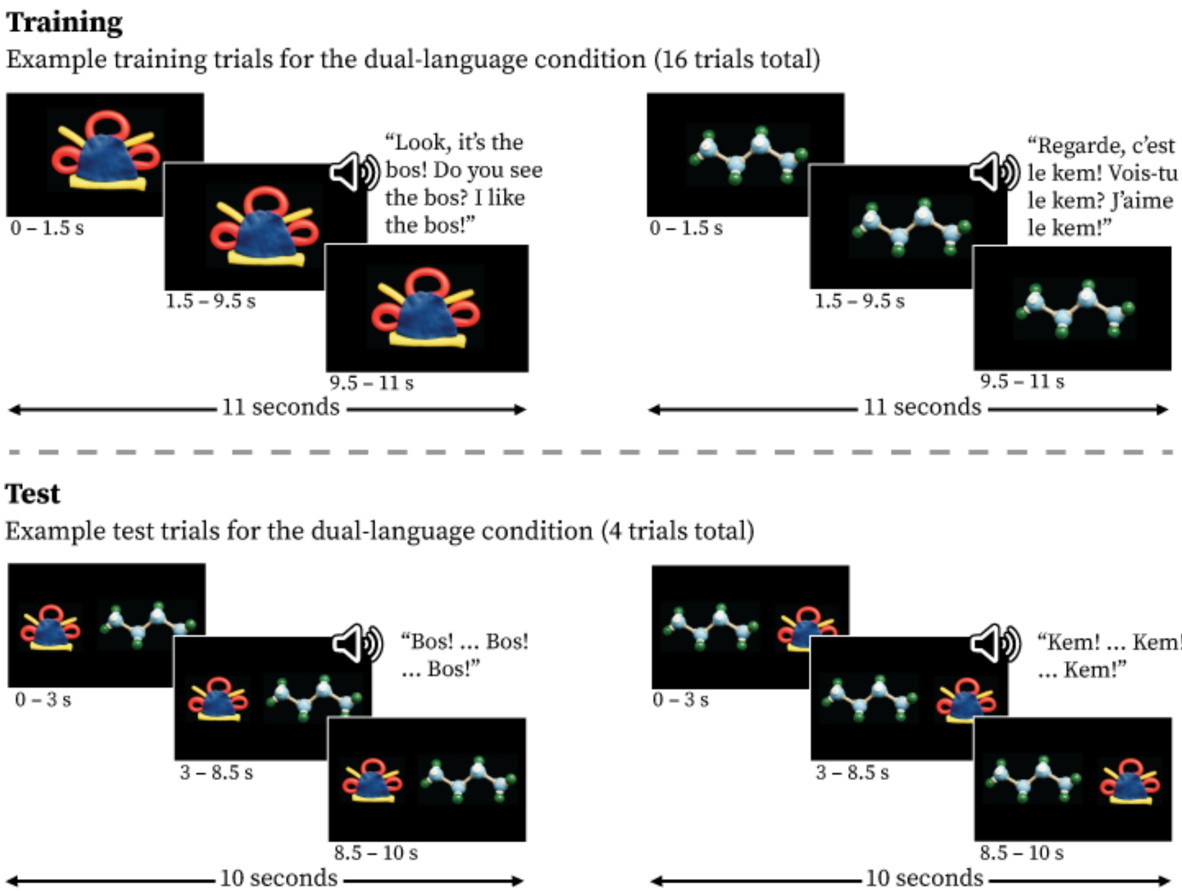

Figure 1. ***Examples of the trial sequence for the training and test phases of the dual-language condition. The single-language condition was identical except all carrier phrases were in the same language.***

Design

Two experimental conditions were developed for the current study: a single-language condition and a dual-language condition. In the single-language condition both objects were labeled in the same language (either English or French) during training. In the dual-language condition, one object was labeled in English and the other object was labeled in French during training. Regardless of condition, each object was labeled 3 times per trial across 8 trials for a total of 24 labeling events per object. Infants thus encountered a total of 16 training trials, presented in one of 8 pseudo-random orders with the constraint that the same word was not encountered for more than two consecutive trials. Orders counterbalanced which word was encountered first, and for the dual-language condition, the pairing of word and language (e.g., whether *bos* was presented in English versus French sentences).

The test phase for all orders and both conditions presented the words *kem* and *bos* in an alternating fashion—never repeating a word or the side the target image appeared on twice in a row. This was to avoid infants developing a side strategy (e.g., if the target appeared on the right twice in a row, infants could anticipate that on the third test trial the target would yet again appear on the right). Test trials were counterbalanced such that for every order where the training phase ended with *kem*, the corresponding test phase started with *bos*, and vice-versa (for test orders see https://osf.io/g6nrv). There were four test trials in all orders, two for *kem* and two for *bos*.

Infants participated in one of 7 different experiments (see Table 1), defined by the experimental condition they completed (single-language versus dual-language) and their own language background. Based on these two factors, we coded a derived variable called familiarity, which related to infants' level of exposure to the sentence frame language and had two possible values: most familiar and least familiar. Note that this variable describes familiarity with the sentence frame languages only—a trial coded as most familiar means the most familiar of English and French, not necessarily the language an infant is most exposed to overall. For example, a Spanish–French bilingual with 70% exposure to Spanish and 30% exposure to French would have French sentence frames coded as 'most familiar' and English sentence frames as 'least familiar', since out of the two sentence frame languages, they have more familiarity with French than English.

In the dual-language condition (Experiments 1–4), infants encountered one word in English and the other in French sentence frames. French–English bilinguals were familiar with both languages, so the word encountered in their dominant language was coded as most familiar, and the one encountered in their non-dominant language was coded as least familiar (Experiment 1). Bilinguals exposed to English or French and another language were familiar with one of the sentence frame languages (either English or French, but not both); in some cases the most familiar sentence frame language was the infants' dominant language (Experiment 2), and in other cases it was the infants' non-dominant language (Experiment 3). As monolinguals were also familiar with only one of the sentence frame languages, this language was coded as most familiar (Experiment 4). In the single-language condition (Experiments 5–7), both novel words were encountered in the same sentence frame language, thus all trials had the same level of familiarity to each infant. The bilinguals tested in the single-language condition were all French–English bilinguals and were purposefully tested with stimuli in their dominant language, thus all sentence frames were coded as most familiar (Experiment 5). Familiarity was coded as most familiar for monolinguals tested with native language sentence frames (Experiment 6), and least familiar for monolinguals tested with sentence frames in the other language (which was foreign to them; Experiment 7).

Under the updated analytic approach, we included percent of exposure to the sentence frame language as a continuous version of the categorical familiarity variable. For example, on trials where the novel word was presented in an English sentence frame, an English monolingual with no exposure to any other language would have an exposure score of 100, a French monolingual with no exposure to English would have a score of 0, a French–English bilingual would have a score of 25 (as one possible value, if they were exposed to English 25% of the time), and a French–Arabic bilingual with no exposure to English would have a score of 0. Thus, higher exposure scores indicate more familiarity with the sentence frame language.

Procedure

A trained research assistant greeted and briefed the parents. Parents then signed the consent form and filled out three questionnaires. The first questionnaire gathered basic demographic information (i.e., infants' general health, birth weight, weeks of gestation and socioeconomic status of the family). The second questionnaire was a detailed interview about the infant's language background starting from birth, using the Language Exposure Questionnaire (LEQ; Bosch & Sebastián-Gallés, 2001) with the Multilingual Approach to Parent Language Estimates (MAPLE; Byers-Heinlein et al., 2019). The third questionnaire (the MacArthur-Bates Communicative Development Inventories: Words and Gestures; Fenson et al., 2007) gathered data on the infant's vocabulary knowledge.

Next, the infant and parent were brought to a sound-attenuated room. The infant sat on the parent's lap in a chair approximately 60 cm away from a Tobii T60 XL eye-tracker, which recorded participants' gaze at 60 Hz. Tobii Studio software was used to display the stimuli on a 24" monitor. Parents were given darkened glasses and headphones playing music, and were instructed not to interact with the child to avoid influencing the infant's responses. Following a 5-point eye-tracking calibration, the experiment started with a 10-second pre-familiarization trial, which consisted of a spinning pinwheel accompanied by a sound. Next, infants saw 16 training trials (8 for *kem* and 8 for *bos*) followed by 4 test trials (2 for *bos* and 2 for *kem*). Between each trial, infants saw an attention-getter (a circle stretching vertically and then horizontally while changing colors) to direct their attention back towards the center of the screen. The experiment ended with the presentation of the spinning pinwheel, and in total lasted approximately 5 minutes.

Results

Traditional Approach

The original experimental design was to conduct a series of individual study conditions with small samples (target n ~ 16) that varied the language(s) of the stimuli (i.e., single-language or dual-language) and the population tested (i.e., monolingual, bilingual English–French, bilingual English/French and another language), and 7 of many possible study conditions (Table 1) were ultimately run.

For the dual-language condition (Experiments 1, 2, 3 and 4), we conducted a preliminary series of paired sample *t*-tests to see if infants preferred the most familiar sentence frame language over the other during training (Table 3). This was to ensure that any differences at test would not be due to differential attention during training. We found no statistically significant differences between groups. However, we found a medium-to-large effect size (Cohen's d = -.64) for bilingual infants dominant in French or English with another second language, who looked longer to training trials in their most familiar language. Given that this was only observed in one of the seven studies and was not statistically significant (even prior to correction for multiple comparisons), this effect is unlikely to be meaningful.

Table 3. ***Total Looking Time in Seconds during Training, Dual-Language Condition***

Language Group	Experiment Number	Most Familiar Language Mean *(SD)*	Least Familiar Language Mean *(SD)*	*t*-test
Bilingual French/English (n = 17)	1	40.27 *(20.88)*	41.72 *(15.94)*	$t(16) = .56$, p = .583, d = .14
Bilingual Dominant in English/ French and L2 Other (n = 10)	2	51.14 *(12.67)*	42.44 *(14.22)*	$t(9) = -2.02$, p = .074, d = -.64
Bilingual Dominant in Other Language and L2 English/French (n = 14)	3	48.23 *(16.35)*	46.75 *(16.03)*	$t(13) = -.71$, p = .490, d = -.19
Monolingual (n = 18)	4	32.65 *(20.25)*	31.69 *(17.41)*	$t(17) = -.34$, p = .739, d = -.08

Note. L2 refers to infants' non-dominant language.

Preliminary analyses also indicated a slight pre-naming preference for looking at the *kem* object in the period of time before the onset of any utterance during the test phase. A *t*-test comparing the proportion looking to each object visible on screen before the onset of the auditory stimulus during test trials (0–3000 ms) showed a statistically significant preference for the *kem* object (*kem M* = .55, *SD* = .15; *bos M* = .45, *SD* = .15), $t(108) = -3.57$, $p = <.001$, $d = -.34$). To account for this difference, we conducted our main analyses using a preference-corrected dependent variable by subtracting each participant's own pre-naming preference for each object from their proportion looking to that target object. This created a variable where a score of zero would indicate no difference between an infant's looking on a given trial and their pre-naming preference for that object, a score greater than zero would indicate more looking to the target object than their pre-naming preference for that object, and a score less than zero would indicate less looking to the target object than their pre-naming preference for that object. Statistical comparisons were then made against zero instead of 50% chance[5.]

Following da Estrela and Byers-Heinlein (2016), only infants with at least one data point for each word (i.e., one for *kem* and one for *bos*) were included in the analyses. A series of *t*-tests revealed that only the bilingual English–French and monolingual infants in the dual-language condition (Experiments 1 and 4) looked at the correct object above chance, but only when the novel word was presented in the least familiar sentence frame language during training (Experiment 1: $M = .07$, $SD = .1$, $t(16) = 2.81$, $p = .012$, $d = .68$; Experiment 4: $M = .15.$, $SD = .21$, $t(17) = 2.94$, $p = .009.$, $d = .69$; see Figures 2 and 3; Table 4). This result was surprising, especially for the monolingual group, given that infants were completely unfamiliar with the sentence frame language. We expected this to be the most challenging context for word learning.

To investigate whether the small sample sizes per group were masking an overall effect, we also performed a *t*-test comparing proportion looking minus infants' pre-naming preference for the target object to zero pooling data from all experiments. This test showed that, on average, infants did look slightly above chance during the test phase ($M = .04$, $SD = .21$), $t(219) = 2.88$, $p = .004$, $d = .19$). Further exploratory analyses suggested that this effect was driven by correct looking to the *bos* object when it was labeled ($M = .05$, $SD = .20$, $t(110) = 2.85$, $p = .005$, $d = .27$), but not the *kem* object ($M = .03$, $SD = .21$), $t(110) = 1.26$, $p = .209$, $d = .12$), above chance levels. Thus, when data were

[5] For transparency, we note that the baseline preference for the *kem* object was discovered during the review process. Earlier versions of the manuscript conducted analyses with comparisons to 50% chance. Results were somewhat similar, except that without the baseline correction we found no evidence from either the traditional or updated analyses that infants learned either of the two words.

pooled, we found possible evidence for learning one of the words, but limited to no evidence for learning the other.

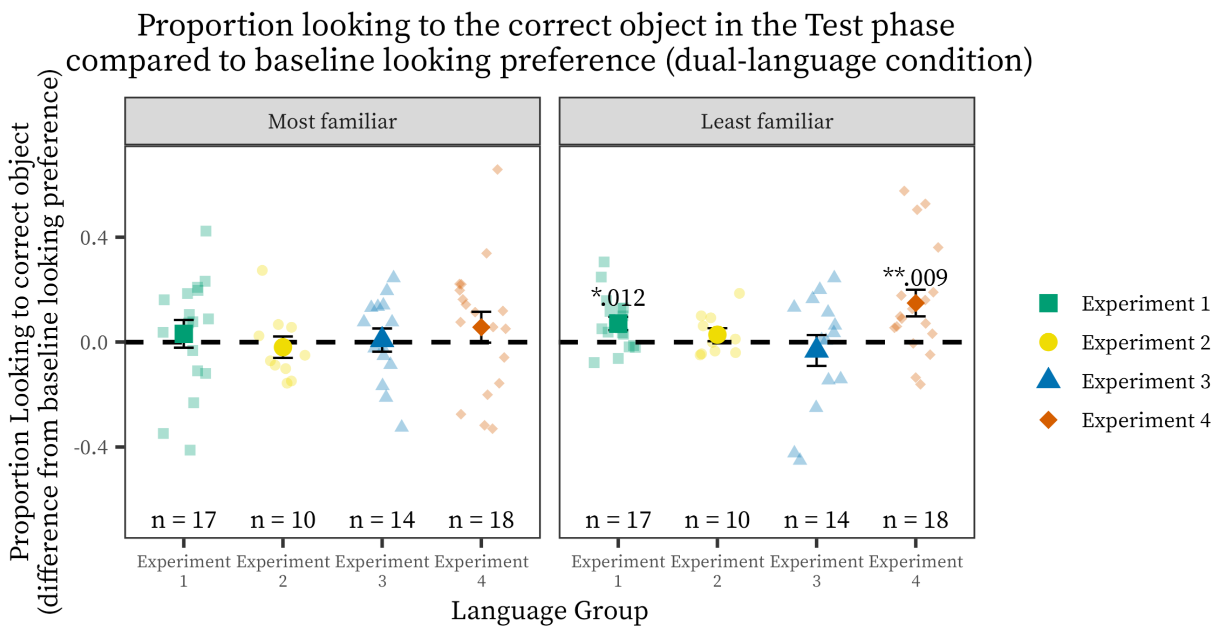

Figure 2. ***Graphs showing proportion looking to the correct object (difference from baseline looking preference) by group in the dual-language condition and standard errors. Same-colour shapes represent an experimental language group. The teal squares represent Experiment 1 (English–French bilinguals). The yellow circles represent Experiment 2 (bilinguals whose first language is English or French with a second language that is not English or French). The blue triangles represent Experiment 3 (bilinguals whose first language is not English or French with English or French as their second language). The orange diamonds represent Experiment 4 (English or French monolinguals). Data are faceted by infants' familiarity with the sentence frame language. Large shapes represent the mean, small shapes represent individual data points, error bars represent the Standard Error, and the dotted line represents no difference from baseline looking preference. The number of participants per mean is indicated with "n =". * p < .05, ** p < .01***

Table 4. ***t-test Results and Means by Group and Condition for the Traditional Analytic Approach***

Language Group	*n*	Sex	Exp. #	Familiarity	Mean	*SD*	*t*	*p*	*df*	*d*
				Dual-Language Condition						
Bilingual English–French	17	7 F	1	Most Familiar (Dominant)	0.03	0.22	0.60	0.559	16	0.14
				Least Familiar (Non-Dominant)	0.07	0.1	2.81	**0.012***	16	0.68
Bilingual Dominant in EN/FR and L2 Other language	10	6 F	2	Most Familiar (Dominant)	-0.02	0.13	-0.48	0.644	9	-0.15
				Least Familiar (Foreign)	0.03	0.08	1.13	0.289	9	0.36
Bilingual Dominant in Other language and L2 EN/FR	14	8 F	3	Most Familiar (Non-Dominant)	0.01	0.16	0.17	0.867	13	0.05
				Least Familiar (Foreign)	-0.03	0.22	-0.54	0.598	13	-0.14
Monolingual	18	8 F	4	Most Familiar (Native)	0.06	0.25	0.96	0.351	17	0.23
				Least Familiar (Foreign)	0.15	0.21	2.94	**0.009****	17	0.69
				Single-Language Condition						
Bilingual EN/FR	19	12 F	5	Most Familiar (Dominant–*Kem*)	0.02	0.15	0.68	0.503	18	0.16
				Most Familiar (Dominant–*Bos*)	0.05	0.17	1.31	0.206	18	0.30
Monolingual	16	9 F	6	Most Familiar (Native–*Kem*)	-0.01	0.36	-0.07	0.944	15	-0.02
				Most Familiar (Native–*Bos*)	0.04	0.28	0.59	0.566	15	0.15
Monolingual	16	7 F	7	Least Familiar (Foreign–*Kem*)	0.06	0.19	1.32	0.206	15	0.33
				Least Familiar (Foreign–*Bos*)	0.05	0.2	0.92	0.371	15	0.23

Note. L2 refers to infants' non-dominant language. F refers to number of females. EN = English, FR = French. Exp. = Experiment number.

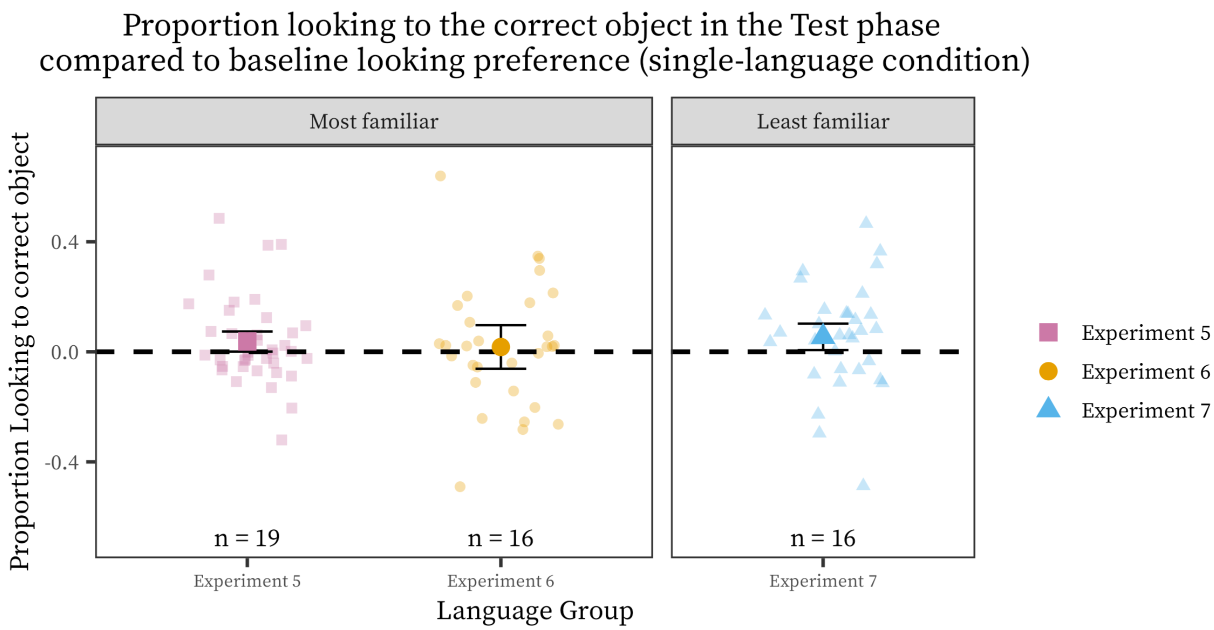

Figure 3. ***Graphs showing proportion looking to the correct object (difference from baseline looking preference) by group in the single-language condition. The purple squares represent Experiment 5 (English–French bilinguals), the orange circles represent Experiment 6 (English or French monolinguals tested in their native language), and the light blue triangles represent Experiment 7 (English or French monolinguals tested in the language they do not know). Large shapes represent the mean, small shapes represent individual data points, error bars represent the Standard Error, and the dotted line represents no difference from baseline looking preference. The number of participants per mean is indicated with "n =". Data are faceted by infants' familiarity with each sentence frame language.***

Updated Approach

Our traditional approach largely tested the performance of small groups of participants against chance level, following the relevant literature at the time the study was designed. More recent discussions on the reproducibility and reliability of psychological science highlight the need for more sensitive analytical approaches that take into consideration the structure of the data (e.g., repeated measures) and that have an appropriate sample size (Bergmann et al., 2018; Oakes, 2017). One well-accepted approach is mixed-effects models (Dixon, 2008). These models have several advantages over traditional methods such as ANOVAs or multiple *t*-tests on different groups. For instance, they can account for the relationship between continuous outcomes (e.g., looking time to the target) and continuous predictors (e.g., language exposure, vocabulary size), which are modeled as fixed effects. They can also account for systematic variability arising from data being grouped (e.g., repeated measures within participants or items), which are modeled as random effects. Furthermore, by modelling fine-grained data (e.g., trial-level data rather than condition averages), these models have greater statistical power and better handling of missing data, even for unbalanced datasets (Baayen et al., 2008; Bates et al., 2015). To harness the richness of our eye-tracking data, we fitted linear mixed-effects models to investigate infant word learning, using the lme4 package for *R* (Bates et al., 2015). All data and scripts are available at https://osf.io/upy7f.

For this analytical approach, we used a larger sample (*N* = 148; see Participants for details). Although larger, we must note that this sample was highly heterogeneous, with infants from diverse linguistic backgrounds (Table 2). We tested whether in this larger sample infants showed word learning and the influence of covariates such as familiarity with sentence-frame language, receptive vocabulary size, and total looking time to objects during the training phase. Given that our Traditional Analysis revealed a preference for *bos* over *kem*, we included target words as a random effect in the model, which would allow us to test the effects of our predictors of interest on word learning while controlling for any differences in looking between the two target objects.

The dependent variable for mixed-models was the proportion of looking time to the labeled object in each trial minus the chance level (.5), so that the intercept would capture overall word learning different from chance. First, we fit an intercept-only model to examine infants' mean accuracy before exploring potential moderators of performance (Table 5). Next, we explored the effects of three continuous variables on learning: the percent of exposure to the sentence frame language, infants' receptive vocabulary size in the sentence frame language, and the total looking time to the objects during the training phase (Table 6). Percentage of exposure to the sentence frame language and vocabulary size allowed us to further explore if or how our participants' language background guided learning. Total looking time to the objects during training allowed us to investigate if participants who were more or less attentive during training would show differences in learning during the test phase. We also ran models on the conditions separately (i.e., one model for the dual-language condition and one for the single-language condition), to see if combining them might be masking some effects. However, there were no additional effects, so these models are not reported here (see Supplemental Materials, Tables S6 to S8, available at https://osf.io/upy7f).

We attempted to fit a maximal random effects structure to our models that included the novel words (*kem* and *bos*) as random slopes and participants as random intercepts (Barr et al., 2013). These models had a singular fit. We then attempted to include the novel words and participants as separate random intercepts. Once again, the models had a singular fit and a closer inspection indicated that there was not enough variability between participants to be included as random intercepts. We then simplified the models to include only the target words as random intercepts. These models converged without a singularity warning and respected the assumptions of normality (see https://osf.io/upy7f for details).

Results are displayed in Tables 5 and 6, and Figure 4. Overall, our reanalysis with this updated approach and the larger sample size confirmed the pattern found in the traditional analyses: there was no evidence of overall word learning while controlling for the difference in looking between *kem* and *bos*, and further, there were no significant relationships between the proportion of looking to the target and (a) exposure to the sentence frame languages, (b) receptive vocabulary size in the sentence frame languages, or (c) the total looking time to the objects during training. Estimates were close to zero for the intercept as well as for all predictors. This means that none of our variables of interest predicted the proportion of infants' looking at the labeled objects (Table 6). Furthermore, our approximate effect size, calculated from the intercept-only model using Brysbaert and Stevens' (2018) approach, was very small (d = 0.09).

Table 5. ***Fixed and Random Effects for the Intercept-Only Model [proportion of looking time - .5 ~ 1 + (1 | target word)]***

Predictors		Estimates	95% Confidence Interval	p
Intercept		0.03	-0.06 – 0.11	0.544
Random Effects				
σ^2				0.06
$\tau_{00\ \text{target word}}$				0.00
ICC				0.05
Observations	476			
Marginal R^2 / Conditional R^2	0.000 / 0.049			

Table 6. ***Fixed and Random Effects for the Pruned Model with Exposure to Sentence Frame Language, Vocabulary size, and Total Looking Time during Training as Predictors of Looking to the Labeled Object [proportion of looking time - .5 ~ exposure + vocabulary + total looking time during training + (1 | target word)]***

Predictors	Estimates	95% Confidence Interval	*p*
Intercept	0.05	-0.06 – 0.16	0.334
Exposure to sentence frame language	-0.00	-0.00 – 0.00	0.557
Receptive vocabulary for sentence frame language	0.00	-0.00 – 0.00	0.409
Total looking time during training	-0.00	-0.00 – 0.00	0.369
Random Effects			
σ^2			0.06
$\tau_{00\ \text{target word}}$			0.00
ICC			0.05
Observations	476		
Marginal R^2 / Conditional R^2	0.003 / 0.052		

To follow up on the finding from the traditional analysis where we found some evidence of learning on *bos* test trials (but not on *kem* test trials), we attempted to fit a model with target word as a fixed effect in addition to our other predictors as fixed effects. Again, models were singular when participants were included as a random effect. Thus, we ran a multiple linear model with these data using the preference difference score as the dependent variable to account for baseline differences in looking toward the two objects. We again found evidence that performance was better for *bos* trials than *kem* trials (see Table 7), after infants' pre-naming baseline looking preferences were accounted for ($\beta 0 = .07$, $p = .031$). However, no other predictors were significant and the model overall explained very little variance in the data ($R^2 = .007$, $F(4,469) = .877$, $p = .477$).

Overall, depending on the model, we found either little evidence of word learning or some evidence of learning one but not both words. Our models also provided little to no account of the observed variance.

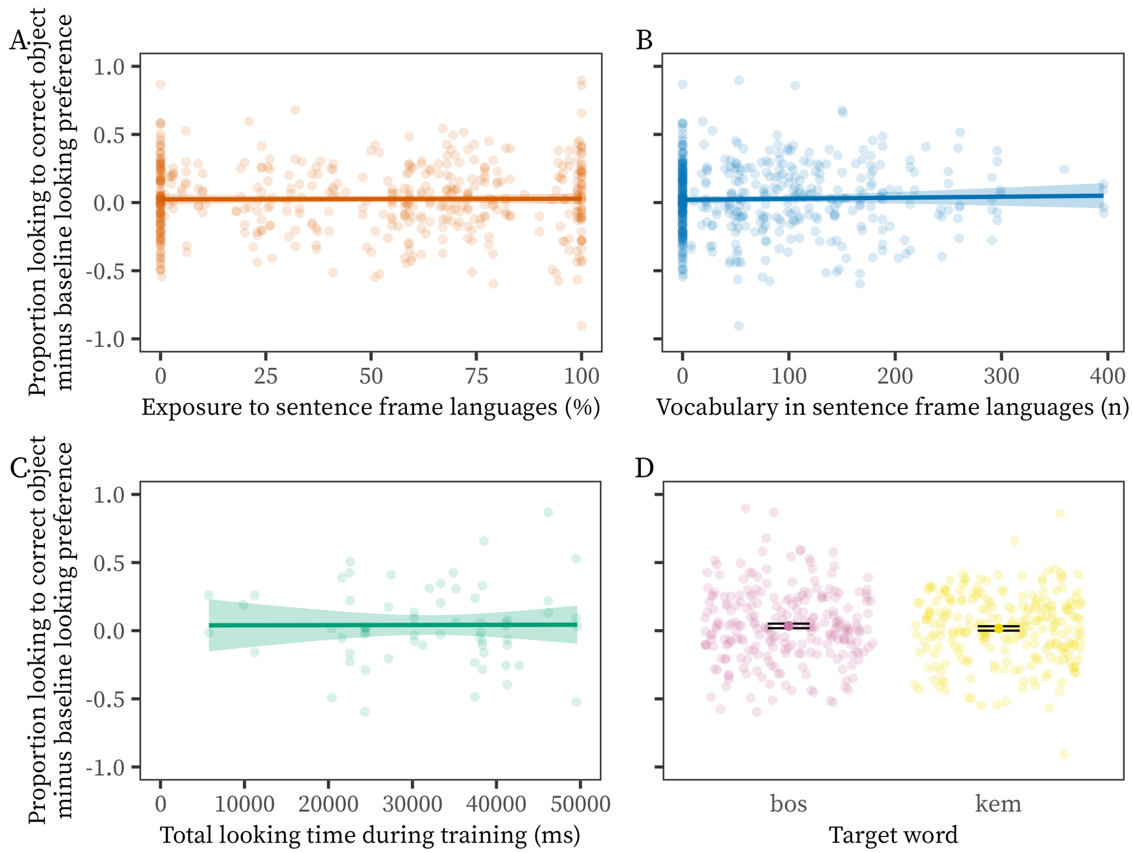

Figure 4. ***Proportion of looking at the correct object as a function of (A) percentage of exposure to the sentence frame languages, (B) vocabulary size (number of words comprehended) in the sentence frame languages, (C) total looking time (ms) during the training phase, and (D) target word. Regression lines, standard errors, and all data points are plotted. Note that chance is 0.***

Table 7. ***Multiple linear regression results using difference score as the criterion***

Predictors		*Estimates*	95% Confidence Interval	*p*
Intercept		0.08	0.01 – 0.15	**0.048**
Target word [*kem*]		-0.02	-0.06 – 0.03	0.439
Exposure to sentence frame language		-0.00	-0.00 – 0.00	0.792
Receptive vocabulary for sentence frame language		0.00	-0.00 – 0.00	0.674
Total looking time during training		-0.00	-0.00 – 0.00	0.165
Observations	476			
R^2 / R^2 adjusted	0.006 / -0.002			

General Discussion

The present study investigated word learning in 14-month-olds from different language backgrounds using a preferential looking paradigm. Following prior research (Fennell & Waxman, 2010; Fernald & Hurtado, 2006; Havy et al., 2016) we assumed that the use of sentence frames would support word learning in infants, and that infants would readily learn the two words that they encountered during the training phase. Moreover, we predicted that language familiarity would play a key role in word learning, with infants showing better learning of word–object associations when they had greater familiarity with the sentence frame language.

First, and quite surprisingly we found only limited evidence for successful word learning in this paradigm. Out of 7 *t*-tests conducted in our traditional approach, only two showed performance that was statistically above chance overall, which we interpret as possible false positives, although given the small samples (n = 10–18/group) false negatives in the other experiments are also possible. Moreover, in our updated approach, which used a larger dataset (N = 148) and had greater statistical power (reducing the chances of both Type I and Type II error), mixed effects models found no evidence of an effect of amount of exposure to the sentence frame language, vocabulary, or attention during the training phase on word learning. By contrast, when data were pooled without including a random effect for item (via *t*-tests and linear regressions), there was some evidence that infants learned one, but not both words. Specifically, when baseline looking preferences were taken into account, there was evidence that infants learned "*bos*" but not "*kem*". We note that successful learning of both of these nonsense words has been previously reported in the literature (Fennell & Byers-Heinlein, 2014; Mattock et al., 2010), making it unlikely that this pattern was driven by our particular choice of stimuli. Overall, evidence for successful word learning in this study was inconsistent.

With respect to familiarity effects, again there was only limited and weak evidence in a direction contrary to hypotheses. Specifically, when traditional analyses were conducted (via separate *t*-tests on data from small groups of infants), two groups of infants showed evidence of learning words presented in frames that were in their least familiar language, but none showed evidence of learning words presented in frames that were in their most familiar language. Again, we note that these analyses had limited statistical power. However, in the updated linear mixed-effects models, which measured familiarity continuously, we did not find an effect of familiarity.

Overall, we believe that the most appropriate interpretation of our results is that word learning in the lab using this paradigm can be challenging for some infants, even with supporting sentence frames. Our findings are unexpected and contrast with previous studies that have reported successful word learning for monolingual 14-month-olds using isolated words (Graf Estes & Hay, 2015; Werker et al., 1998; Yin & Csibra, 2015) and sentence frames (da Estrela & Byers-Heinlein, 2016; Fennell & Waxman, 2010). Importantly, our task was designed to be easy and conducive to word learning. To this end, we used sentence frames which were meant to provide further linguistic cues and presented the target words in a sentence-final position to increase their salience (e.g., Fennell & Byers-Heinlein, 2014; Fernald & Hurtado, 2006). In addition, each word was repeated multiple times during training (3 times per trial for 8 trials, for a total of 24 exposures to each word–object pairing) and we taught infants only two novel words to reduce their cognitive load. Even so, neither monolingual nor bilingual infants showed evidence of learning both words, even the word–object pairs presented in the sentence frame language that was most familiar to them.

Although our experiment was designed to provide a facilitative word learning opportunity for infants, it is possible that the task was simply too taxing. We used consistent word–object pairings that have been used successfully in previous studies of word learning (Werker et al. 1998; Werker et al., 2002; Fennell et al., 2007), but it is possible that these stimuli were suboptimal[6]. One crucial difference between our study

[6] For example, the pairings might have violated sound symbolic associations (e.g., Sidhu & Pexman, 2018). We thank an anonymous reviewer for raising this point.

and previous studies that have successfully shown word learning with 14-month-olds (e.g., Werker et al., 1998) is that our study presented infants with a fixed number of training trials rather than presenting training trials according to a habituation criterion (as in the Switch task), which may make our task less effective as it did not adapt to each infant's learning (Yoshida et al., 2009). It could also be the case that infants required additional familiarization with the task structure (e.g., familiar-word trials presented before training, where a known word is associated with a known object to cue the task, see Fennell & Waxman, 2010 and May & Werker, 2014). However, this interpretation contrasts with reports in the published literature. For example, Schafer and Plunkett (1998) reported successful word learning after 12 presentations of each of 2 novel word–referent pairs in 15-month-olds using a similar paradigm to that implemented in our study (though they also presented familiar-word trials between the novel word trials). It is also possible that, rather than presenting infants with too few training trials, we presented them with too many, ultimately leading to boredom and disengagement from the task. This interpretation is supported by the high levels of attrition we observed in our task, a point that we return later in this section. Overall, the optimal amount of exposure to novel words in lab word learning tasks remains unclear.

It is also possible that sentence frames made our task more challenging, contrary to our intentions. We used sentence frames following prior research with monolingual and bilingual infants showing that they have a facilitative effect (e.g., Fennell & Waxman, 2010 in 14-month-olds; Fennell & Byers-Heinlein, 2014 in 17-month-olds; Fernald & Hurtado, 2006 in 18-month-olds). Thus, we expected that sentence frames would support word learning, particularly for bilingual infants, since this additional information might help them identify the language in which a novel word is presented. Yet, this did not appear to be the case. Similarly, it is possible that using isolated words during testing might have made the task more challenging, since during training sentence frames were used. Future studies could compare experimental conditions that vary on the use of isolated words versus sentence frames (e.g., Morini & Newman, 2019), to disentangle the effect that additional linguistic information has on early word learning.

Another possible explanation is that infants did successfully learn both words presented during training trials, but our test phase was not sufficiently sensitive to detect this learning. It could be that the 4 test trials included in our study (2 per novel word) were not enough to robustly detect learning, especially because some infants did not provide valid data for both words during the test phase. Prior studies using a preferential looking paradigm reported successful word learning when infants were tested with 4–8 novel word test trials per condition (e.g., Chen et al., 2020; Schafer & Plunkett, 1998; Tan & Schafer, 2005; Yoshida et al., 2009), although studies using the Switch word learning paradigm have often used only two test trials (see data compiled by Tsui et al., 2019). Increasing the number of test trials per infant might increase the chances of capturing learning in this hard-to-test population, and would most likely generate a better representation of infants' true response to the task, thus decreasing noise and increasing statistical power (DeBolt et al., 2020).

Moreover, we selected the preferential looking task based on extant literature suggesting that it might be more sensitive to detect word learning than other paradigms such as the Switch task (Yoshida et al., 2009). However, many studies reporting successful word learning in infants have used the Switch task (see Tsui et al., 2019 for a meta-analysis), and it may be that the Switch task is in fact more sensitive, or at least more forgiving when infants have only learned one of two words. In the Switch task at least two novel words are paired with two referents (word A with object A, word B with object B). At test, some trials show the label and referent that were previously paired (A with A; Same trials) and some trials show a label with the other referent (A with B; Switch trials). In this paradigm, infants only need to associate one word–referent pair to recognize a word–object violation. If infants learn that word A should be associated with object A, they should be able to detect the violation when word A is paired with object B. However, in our preferential looking paradigm, infants had to correctly identify both word–object pairings to show learning of each word. Moreover, it may be that detecting a pairing violation (dishabituating in the Switch task) can potentially be accomplished with weaker knowledge than looking towards a correct referent in a preferential looking paradigm. Tsui et al.'s (2019) meta-analysis reported an average effect of Cohen's d = 0.33, 95% CI [0.03, 0.63] in comparable studies using the Switch task (i.e., 14-month-olds learning dissimilar-sounding words), which was moderate and much larger than the approximate d = 0.09 we observed in our own data (Brysbaert & Stevens, 2018). Nonetheless, little work has compared infants' performance in the Switch task to a preferential looking test using the same learning task (although see Yoshida et al., 2009), and thus it remains an open methodological question which tasks are most sensitive for testing infant word learning. Developing maximally sensitive and reliable tasks should be a priority for research on infant word learning.

Another well-documented possibility is that sampling and measurement error in the context of small samples can lead to highly variable, and unreliable, effect-size estimates (Brysbaert, 2021; Lindsay, 2020; Oakes, 2017). For instance, underpowered studies can lead to exaggerated effect size estimates that, combined with publication bias favouring positive results, might end up published, whereas null results with similar sample sizes end up in the file drawer (Rosenthal, 1979). As mentioned in the Introduction, our per group sample size was chosen back in 2012, following sample sizes from other studies in the field (e.g., Fennell & Werker, 2003; Mattock et al., 2010), and after 7 years of testing infants, we were not able to achieve our (small) target sample in all groups. In retrospect, we acknowledge that our original experimental plan was both overly ambitious and underpowered. Even when these small groups were combined in our updated approach, the sample was very heterogeneous, limiting our explanatory scope. At the same time, given our large overall sample, we would have expected to find statistically reliable learning of both words, even if there were some moderators of an overall positive effect size. However, our mixed effects models explored three different variables – percent exposure to the sentence frame language, receptive vocabulary in the sentence frame language, and attention during training – and found no effects (estimated effect size of d = 0.09). In fact, it was surprising that neither percentage exposure nor vocabulary size modulated performance in this task, given prior studies reporting the influence of these variables in word learning (e.g., Bion et al., 2013; Werker et al., 2002).

Despite these unexpected and mixed results, we believe that there is value in sharing our study, as it shows some of the drawbacks to using traditional methodologies and conventional sample sizes. Open science practices centered around transparency and collaboration, combined with more advanced statistical analyses, have an enormous potential to inform future studies on infant word learning. By planning adequate sample sizes (using a-priori power analyses and simulations), pre-registering analytical pipelines, and sharing materials, data, and research reports, we can work toward more reliable findings in the field. For instance, readers can use our openly shared materials, data, and analysis scripts (open repository: https://osf.io/upy7f/) to both reproduce our methodological and analytical decisions and build on them when designing future investigations on the topic.

Another important issue our study faced is the reduction of our initial sample size. Though we tested 288 14-month-old infants, after implementing our exclusion criteria we lost 62% of our participants for the traditional approach and 49% for the updated one. A large proportion of our exclusions (23% for the traditional approach) were related to infants' language background, which can be a particular challenge of studies with bilingual populations. Within the other excluded infants, the largest reason for exclusion was fussiness and inattention (15%), a major issue in infant research. In our updated analytical approach, we were able to include 38 additional participants who had been excluded using the traditional analytic approach. Including these additional infants did not change the pattern of results that we observed. Moreover, our attrition rates, while high, are within the range reported in previous studies including infants of similar ages (e.g., 26% exclusion rate in Experiment 1 and 32% in Experiment 2 in Graf Estes et al., 2007; 44% exclusion rate in Yu & Smith, 2011, 35% exclusion rate in Escudero &

Kalashnikova, 2020). While high attrition can reduce power, our sample size was still large overall.

One way to achieve larger sample sizes and more robust results is with collaborations between different research labs. When participants are recruited in multiple locations, it is easier to obtain larger samples, and the results are also more generalizable. Although some researchers may find it more challenging than others to conduct large studies on their own or to engage in large-scale collaborations, it is important to consider the value of carrying out research that may not be sufficiently powered in the first place (Brysbaert, 2021; Oakes, 2017). In recent years, more opportunities to take part in such large-scale collaborations have become available, and often do not require extensive resources to join (e.g., Byers-Heinlein et al., 2020; Frank et al., 2017; ManyBabies Consortium, 2020). Similarly, open science practices such as open sharing of stimuli and protocols between researchers can be useful in identifying procedures and materials (e.g., novel objects, number of trials) that are better tolerated by infants at specific ages, reducing fussiness and participant loss.

There are several other explanations for weak or null results that we also considered, but found unlikely[7]. First, it has been proposed that null results in some infant looking time studies could be due to some infants showing a familiarity preference, and others a novelty preference, which could average out to a null result (e.g., DePaolis et al., 2016). However, this line of reasoning does not clearly apply to preferential looking paradigms like ours, where infants are always expected to look towards the labeled target rather than the distractor object. Second, one might ask whether incidental factors such as the room where data were collected, or the particular speaker who recorded our stimuli, contributed to our null results. Our lab has conducted many other studies with positive results in the same space, and using similar procedures for recording stimuli, training research assistants, and testing infants, making it unlikely that these factors would affect this study in particular. In the future, Big Team Science efforts such as ManyBabies might provide insight into whether and how such incidental sources of variation relate to effect sizes (Frank et al., 2017). Third, it is possible that there is an error in our data analysis pipeline. However, this seems unlikely as looking time was gathered via an eyetracker, and the analysis was fully automated in *R* and was double checked. We have provided our materials, raw data, and analysis code on the Open Science Framework such that it can be checked or even further analysed by other researchers, who might come to different conclusions or identify limitations that we did not. We would welcome this type of feedback.

[7] We acknowledge the peer-review process for raising these possibilities.

Regardless of any potential limitations of the current experiment as designed and performed, our results are nonetheless surprising. There is a vast body of published research showing successful word learning in the lab with infants, even with methods and small sample sizes comparable to ours. Taking into account the publication bias for positive results (Carter et al., 2019; Rosenthal, 1979), it is impossible to know how many "unsuccessful" infant word learning studies languish in the file drawer. If they do exist in significant numbers, their absence from the literature may distort the picture of how easily infants learn new words in the lab, and, by consequence, any generalization to the real world outside the lab. Increasingly, journals, editors, and reviewers are recognizing the importance of publishing null results, and researchers are embracing open science practices such as pre-registration and registered reports (Tsuji et al., 2020a). With these efforts, the published literature might present a more accurate picture of the true effects in hard-to-test populations, like young infants. Additionally, developing large-scale collaborations across labs, with greater power and sample diversity, might also contribute to a better characterization of infants' word learning abilities.

Overall, our study raises the possibility that word learning in the lab could in some cases be challenging for 14-month-old bilinguals and monolinguals, despite the presence of sentence frames that could support learning. The case study presented here highlights the need for and value of open science practices to advance our understanding of infant development.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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

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

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives, 14*(3), 142–149. https://doi.org/10.1111/cdep.12373

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, 109*(9), 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*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition, 126*(1), 39-53. https://doi.org/10.1016/j.cognition.2012.08.008

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research. *Quarterly Journal of Experimental Psychology, 73*(1), 1–19. https://doi.org/10.1177/1747021819886519

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy, 2*(1), 29–49. https://doi.org/10.1207/S15327078IN0201_3

Brysbaert, M. (2021). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition, 24*(5), 813-818. https://doi.org/10.1017/S1366728920000437

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1), 9. https://doi.org/10.5334/joc.10

Burnham, D., Singh, L., Mattock, K., Woo, P. J., & Kalashnikova, M. (2018). Constraints on tone sensitivity in novel word learning by monolingual and bilingual infants: tone properties are more influential than tone familiarity. *Frontiers in Psychology, 8*, 2190. https://doi.org/10.3389/fpsyg.2017.02190

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children’s vocabulary size. *Bilingualism: Language and Cognition, 16*(1), 32–48. https://doi.org/10.1017/S1366728912000120

Byers-Heinlein, K. (2015). Methods for studying infant bilingualism. In J. W. Schwieter (Ed.), *The Cambridge Handbook of Bilingual Processing* (pp. 133–154). Cambridge University Press. https://doi.org/10.1017/CBO9781107447257.005

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne, 61*(4), 349–363. https://doi.org/10.1037/cap0000216

Byers-Heinlein, K., Fennell, C. T., & Werker, J. F. (2013). The development of associative word learning in monolingual and bilingual infants. *Bilingualism: Language and Cognition, 16*(1), 198–205. https://doi.org/10.1017/S1366728912000417

Byers-Heinlein, K., Jardak, A., Fourakis, E., & Lew-Williams, C. (2022). Effects of language mixing on bilingual children's word learning. *Bilingualism: Language and Cognition, 25*(1), 55-69. doi:10.1017/S1366728921000699

Byers-Heinlein, K., Schott, E., Gonzalez-Barrero, A. M., Brouillard, M., Dubé, D., Jardak, A., Laoun-Rubenstein, A., Mastroberardino, M., Morin-Lessard, E., Pour Iliaei, S., Salama-Siroishka, N., & Tamayo, M. P. (2019). MAPLE: A Multilingual Approach to Parent Language Estimates. *Bilingualism: Language and Cognition,* 1–7. https://doi.org/10.1017/S1366728919000282

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., ... & Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science, 4*(1), 1-30. https://doi.org/10.1177/2515245920974622

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: Infants’ language experience influences the development of a word-learning heuristic. *Developmental Science, 12*(5), 815–823. https://doi.org/10.1111/j.1467-7687.2009.00902.x

Canfield, R. L., Smith, E. G., Brezsnyak, M. P., Snow, K. L., Aslin, R. N., Haith, M. M., Wass, T. S., & Adler, S. A. (1997). Information processing through the first year of life: A longitudinal study using the visual expectation paradigm. *Monographs of the Society for Research in Child Development, 62*(2), 1-160. https://doi.org/10.1111/j.1540-5834.1997.tb00511.x

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, *15,* 17–29.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/doi.org/10.1177/2515245919847196

Chen, H., Labertonière, D., Cheung, H., & Nazzi, T. (2020). Infant learning of words in a typologically distant nonnative language. *Journal of Child Language*, *47*(6), 1276–1287. https://doi.org/10.1017/S0305000920000161

Curtin, S., Fennell, C. T., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word-object associative learning. *Developmental Science, 12*(5), 725–731. https://doi.org/10.1111/j.1467-7687.2009.00814.x

da Estrela, C., & Byers-Heinlein, K. (2016). Vois-tu le kem? Do you see the bos? Foreign word learning at 14 months. *Infancy*, *21*(4), 505–521. https://doi.org/10.1111/infa.12126

Dal Ben, R., de Hollanda Souza, D., & Hay, J. F. (2019). Cross-situational word learning: Systematic review and meta-analysis [Manuscript in preparation]. https://doi.org/10.17605/OSF.IO/GU9RB

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy*, *25*(4), 393–419. https://doi.org/10.1111/infa.12337

DePaolis, R. A., Keren-Portnoy, T., & Vihman, M. (2016). Making sense of infant familiarity and novelty responses to words at lexical onset. *Frontiers in Psychology, 7*, 715. https://doi.org/10.3389/fpsyg.2016.00715

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Echols, C. H., & Newport, E. L. (1992). The role of stress and position in determining first words. *Language Acquisition, 2*(3), 189–220. https://doi.org/10.1207/s15327817la0203_1

Escudero, P., & Kalashnikova, M. (2020). Infants use phonetic detail in speech perception and word learning when detail is easy to perceive. *Journal of Experimental Child Psychology, 190,* 104714. https://doi.org/10.1016/j.jecp.2019.104714

Fennell, C. T., & Byers-Heinlein, K. (2014). You sound like Mommy: Bilingual and monolingual infants learn words best from speakers typical of their language environments. *International Journal of Behavioral Development, 38*(4), 309–316. https://doi.org/10.1177/0165025414530631

Fennell, C. T., Byers-Heinlein, K., & Werker, J. F. (2007). Using speech sounds to guide word learning: The case of bilingual infants. *Child Development, 78*(5), 1510–1525. https://doi.org/10.1111/j.1467-8624.2007.01080.x

Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development, 81*(5), 1376–1383. https://doi.org/10.1111/j.1467-8624.2010.01479.x

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech, 46*(2–3), 245–264. https://doi.org/10.1177/00238309030460020901

Fenson L., Marchman V. A., Thal D. J., Dale P. S., Reznick S., Bates E. (2007). *MacArthur-Bates communicative development inventories* (2nd ed.). Brookes Publishing.

Fernald, A., & Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental Science, 9*(3), F33–F40. https://doi.org/10.1111/j.1467-7687.2006.00482.x

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project.* MIT Press.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421–435. https://doi.org/10.1111/infa.12182

Gonzalez-Gomez, N., Poltrock, S., & Nazzi, T. (2013). A “bat” is easier to learn than a “tab”: Effects of relative phonotactic frequency on infant word learning. *PLoS One, 8*(3), e59601. https://doi.org/10.1371/journal.pone.0059601

Graf Estes, K. (2014). Learning builds on learning: Infants’ use of native language sound patterns to learn words. *Journal of Experimental Child Psychology, 126,* 313–327. https://doi.org/10.1016/j.jecp.2014.05.006

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*(3), 254-260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Graf Estes, K., & Hay, J. F. (2015). Flexibility in bilingual infants’ word learning. *Child Development, 86*(5), 1371–1385. https://doi.org/10.1111/cdev.12392

Havy, M., Bouchon, C., & Nazzi, T. (2016). Phonetic processing when learning words: The case of bilingual infants. *International Journal of Behavioral Development, 40*(1), 41–52. https://doi.org/10.1177/0165025415570646

Hirsh-Pasek, K., Golinkoff, R. M., & Hollich, G. (2000). An emergentist coalition model for word learning: Mapping words to objects is a product of the interaction of multiple cues. In R. M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L. Smith, A. L. Woodward, N. Akhtar, M. Tomasello, & G. Hollich (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 136–164). Oxford University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kalashnikova, M., Escudero, P., & Kidd, E. (2018). The development of fast-mapping and novel word retention strategies in monolingual and bilingual infants. *Developmental Science, 21*(6), e12674. https://doi.org/10.1111/desc.12674

Kalashnikova, M., Mattock, K., & Monaghan, P. (2015). The effects of linguistic experience on the flexible use of mutual exclusivity in word learning. *Bilingualism: Language and Cognition, 18*(4), 626-638. https://doi.org/10.1017/S1366728914000364

Kandhadai, P., Hall, D. G., & Werker, J. F. (2017). Second label learning in bilingual and monolingual infants. *Developmental Science, 20*(1), e12429. https://doi.org/10.1111/desc.12429

Kannass, K. N., & Oakes, L. M. (2008). The development of attention and its relations to language in infancy and toddlerhood. *Journal of Cognition and Development, 9*(2), 222–246. https://doi. org/10.1080/15248370802022696

Kremin, L. V., & Byers-Heinlein, K. (2021). Why not both? Rethinking categorical and continuous approaches to bilingualism. *International Journal of Bilingualism, 25*(6), 1560-1575. https://doi.org/10.31234/osf.io/nkvap

Kremin, L. V., Alves, J., Orena, A. J., Polka, L., & Byers-Heinlein, K. (2022). Code-switching in parents' everyday speech to bilingual infants. *Journal of Child Language, 49*(4), 714-740. https://doi.org/10.1017/S0305000921000118

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2015). Slowing down fast mapping: Redefining the dynamics of word learning. *Child Development Perspectives, 9*(2), 74–78. https://doi.org/10.1111/cdep.12110

Lany, J. (2014). Judging words by their covers and the company they keep: Probabilistic cues support word learning. *Child Development, 85*(4), 1727-1739. https://doi.org/10.1111/cdev.12199

Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology/Psychologie Canadienne, 61*(4), 310–317. https://doi.org/10.1037/cap0000222

Luk, G. (2015). Who are the bilinguals (and monolinguals)? *Bilingualism: Language and Cognition, 18*(1), 35–36. https://doi.org/10.1017/S1366728914000625

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology, 25*(5), 605–621. https://doi.org/10.1080/20445911.2013.795574

Mattock, K., Polka, L., Rvachew, S., & Krehm, M. (2010). The first steps in word learning are easier when the shoes fit: Comparing monolingual and bilingual infants: Phonetic variability and word learning. *Developmental Science, 13*(1), 229–243. https://doi.org/10.1111/j.1467-7687.2009.00891.x

May, L., & Werker, J. F. (2014). Can a click be a word?: Infants' learning of non-native words. *Infancy, 19*(3), 281-300. https://doi.org/10.1111/infa.12048

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review, 119*(4), 831–877. https://psycnet.apa.org/doi/10.1037/a0029872

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Morini, G., & Newman, R. S. (2019). Dónde está la ball? Examining the effect of code switching on bilingual children's word recognition. *Journal of Child Language, 46*(6), 1238–1248. https://doi.org/10.1017/S0305000919000400

Oakes, L. M. (2017). Sample Size, statistical power, and false conclusions in infant looking-time research. *Infancy, 22*(4), 436–469. https://doi.org/10.1111/infa.12186

Place, S., & Hoff, E. (2011). Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child Development, 82*(6), 1834–1849. https://doi.org/10.1111/j.1467-8624.2011.01660.x

R Core Team (2020). *R:* A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology, 59*(1), 96-121. https://doi.org/10.1016/j.cogpsych.2009.02.002

Rocha-Hidalgo, J., & Barr, R. (2022). Defining bilingualism in infancy and toddlerhood: A scoping review. *International Journal of Bilingualism.* https://doi.org/10.1177/13670069211069067

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641.

Schafer, G., & Plunkett, K. (1998). Rapid word learning by fifteen-month-olds under tightly controlled conditions. *Child Development, 69*(2), 309–320. https://doi.org/10.1111/j.1467-8624.1998.tb06190.x

Schott, E., Kremin, L. V., & Byers-Heinlein, K. (2022). The youngest bilingual Canadians: Insights from the 2016 census regarding children aged 0–9 years. *Canadian Public Policy, 48*(2), 254-266. https://doi.org/10.3138/cpp.2021-064

Schott, E., Rhemtulla, M., & Byers-Heinlein, K. (2019). Should I test more babies? Solutions for transparent data peeking. *Infant Behavior and Development, 54*, 166–176. https://doi.org/10.1016/j.infbeh.2018.09.010

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences, 108*(15), 6038–6043. https://doi.org/10.1073/pnas.1017617108

Sidhu, D.M., & Pexman, P.M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review, 25*, 1619–1643. https://doi.org/10.3758/s13423-017-1361-1

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singh, L., Fu, C. S. L., Tay, Z. W., & Golinkoff, R. M. (2018). Novel word learning in bilingual and monolingual infants: Evidence for a bilingual advantage. *Child Development, 89*(3), e183–e198. https://doi.org/10.1111/cdev.12747

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*(6640), 381–382. https://doi.org/10.1038/41102

Tan, S. H., & Schafer, G. (2005). Toddlers' novel word learning: Effects of phonological representation, vocabulary size and parents' ostensive behaviour. *First Language, 25*(2), 131–155. https://doi.org/10.1177/0142723705050338

Taxitari, L., Twomey, K. E., Westermann, G., & Mani, N. (2020). The limits of infants' early word learning. *Language Learning and Development, 16*(1), 1-21.

https://doi.org/10.1080/15475441.2019.1670184

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology, 55*(5), 934–950. https://doi.org/10.1037/dev0000699

Tsuji, S., Cristia, A., Frank, M. C., & Bergmann, C. (2020a). Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development. *Zeitschrift Für Psychologie, 228*(1), 50–61. https://doi.org/10.1027/2151-2604/a000393

Tsuji, S., Jincho, N., Mazuka, R., & Cristia, A. (2020b). Communicative cues in the absence of a human interaction partner enhance 12-month-old infants' word learning. *Journal of Experimental Child Psychology,* 191, 104740. https://doi.org/10.1016/j.jecp.2019.104740

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology, 34*(6), 1289–1309. https://psycnet.apa.org/doi/10.1037/0012-1649.34.6.1289

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*(2), 197–234. https://doi.org/10.1080/15475441.2005.9684216

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3*(1), 1-30. https://doi.org/10.1207/S15327078IN0301_1

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology, 30*(4), 553–563. https://psycnet.apa.org/doi/10.1037/0012-1649.30.4.553

Yin, J., & Csibra, G. (2015). Concept-based word learning in human infants. *Psychological Science, 26*(8), 1316-1324. https://doi.org/10.1177/0956797615588753

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science, 12*(3), 412–418. https://doi.org/10.1111/j.1467-7687.2008.00789.x.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science, 14*(2), 165-180. https://doi.org/10.1111/j.1467-7687.2010.00958.x

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. Cognition*, 125*(2), 244-262. https://doi.org/10.1016/j.cognition.2012.06.016

Data, code and materials availability statement

All data, code and stimuli used in the present study are available at the Open Science Framework. Stimuli are available at https://osf.io/g6nrv. Data and scripts are available at https://osf.io/upy7f.

Ethics statement

The current study was conducted according to the Declaration of Helsinki and ethics approval was obtained from the Human Research Ethics Board of Concordia University (certification numbers UH2011-041 and 10000439). All participants' parents gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

AMGB: Writing - Original Draft, Writing - Review & Editing, Formal Analysis. RDB: Writing - Review & Editing, Formal Analysis. HK: Writing - Review & Editing, Formal Analysis. KBH: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

We are grateful to all the families who participated in this research. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada to KBH (402470-2011 and 2018-04390), a FQRSC postdoctoral fellowship to AMGB (2018-B3-205717), and a Concordia Horizon postdoctoral fellowship to RDB.

License