WILEY

INFANCY THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES

RESEARCH ARTICLE OPEN ACCESS

Infants' Knowledge of Individual Words: Investigating Links Between Parent Report and Looking Time

Melanie López Pérez | Charlotte Moore | Andrea Sander-Montant | Krista Byers-Heinlein

Concordia University, Montréal, Canada

Correspondence: Melanie López Pérez (M_OPEZPE@live.concordia.ca)

Received: 2 May 2024 | **Revised:** 1 October 2024 | **Accepted:** 9 November 2024

Keywords: bilingualism | CDI | individual word knowledge | looking-while-listening | vocabulary development

ABSTRACT

Assessing early vocabulary development commonly involves parent report methods and behavioral tasks like looking-while-listening. While both yield reliable aggregate scores, findings are mixed regarding their reliability in measuring infants' knowledge of individual words. Using archival data from 126 monolingual and bilingual 14–31-month-olds, we further examined links across these methods at the word level, while controlling for potentially confounding child-level factors. When data were averaged at the child level, performance on the looking-while-listening task correlated well with parent-reported word production of the same words, as expected. However, mixed-effects model comparisons suggested that at the word level, looking-while-listening performance was significantly predicted by age and total productive vocabulary, but not by parent-reported knowledge of a word once these factors were controlled for. These findings invite careful consideration regarding the adequacy of these two popular methods for capturing children's idiosyncratic knowledge of individual words.

1 | Introduction

Children are remarkable language learners, showing rapid early vocabulary development. This development is commonly assessed via parent reports and behavioral measures. These measures are designed to estimate children's vocabulary knowledge in aggregate, but researchers also sometimes use these methods to extract word-level information, like whether a child can understand/say a specific word like "ball." The reliability of such data can be assessed by linking parent-reported word knowledge (i.e., whether infants are reported to say the word "ball") to behaviorally-measured word knowledge (i.e., whether infants look at an image of a ball when labeled), but studies to date have yielded mixed results (Houston-Price, Mather, and Sakkalou 2007; Poulin-Dubois et al. 2013; Styles and Plunkett 2009; Yoder, Warren, and Biggar 1997). Thus, the current study used data from a large sample of infants to investigate whether any word-level relationship holds between the two foremost current methods—parent reports and behavioral measures—when also statistically accounting for child-level factors known to play a role in word learning like age, language experience, and total vocabulary size. In the sections below, we first discuss how each type of measure is used and their ability to assess children's knowledge of individual words. We then review child-level factors associated with children's vocabulary knowledge and discuss how they could potentially affect what is being captured by item-level relationships between parent reports and behavioral measures.

1.1 | Measures of Infants' and Toddlers' Vocabulary

Over the years, different measures have been created to evaluate young children's vocabulary knowledge. However, as for most psychological measures, capturing this knowledge can be challenging due to its latent nature. Word comprehension is particularly difficult to assess, as it must be inferred from other observable behaviors. Similarly, interpreting early spoken words can be difficult, because early vocalizations are not yet adult-like in form (Vihman and McCune 1994). For example, it can be challenging to deem a vocalization like "ba" as "ball," "baby," or a meaningless babble. While no gold standard exists for measuring word knowledge (a term we used here broadly to refer to word comprehension and/or word production), two of the most widely used methods for assessing word knowledge are parent report measures, like the MacArthur-Bates Communicative Development Inventories (CDI; Fenson et al. 1993) and gaze-based measures such as the looking-while-listening paradigm (Fernald et al. 2008).

1.1.1 | Parent Report: CDI

The CDI is a well-known questionnaire that researchers and clinicians use to measure young children's vocabulary knowledge. Parents indicate which of several hundred early-acquired words their children understand (for younger children only) and/or say (for both younger and older children) using a binary response (no/yes scored as 0/1), with adaptations available for different languages (Dale and Penfold 2011). Customarily, word-level scores are summed to create an aggregated total vocabulary score, which serves as a proxy for children's receptive or expressive vocabulary size. Previous validations have shown moderate-to-strong correlations between children's total CDI vocabulary score and other direct language measures like language samples, real-object naming tasks, and behavioural measures like looking-time tasks (Fenson et al. 1994; for a review see J. Law and Roy 2008). Furthermore, the CDI has been validated across diverse populations, including monolingual (e.g., Fenson et al. 1993) and bilingual children (Marchman and Martínez-Sussmann 2002; Pearson, Fernández, and Oller 1993). For bilinguals, the CDI has been adapted and used across a variety of languages and language-pairs both cross-sectionally and longitudinally, overall with good results (for a scoping review, see Weisleder et al. 2024).

1.1.2 | Behavioral Measures: Looking-While-Listening

Direct behavioral assessments of children's early vocabulary are also valuable, especially for children who produce little language. A common behavioral task for measuring word comprehension is the looking-while-listening task (Fernald et al. 2008), with similar variants like the intermodal preferential looking procedure (Hirsh-Pasek and Golinkoff 1996) wherein children are shown a target and distractor image on screen (e.g., a boat and a shoe), while hearing the label for a target object (e.g., "Look at the shoe!"), and their eye gaze is recorded. Gaze data provides several measures of word comprehension, like accuracy and reaction time, typically aggregated for each child, and within each condition if the study involves a manipulation. Commonly, researchers calculate the proportion of looking time to the target object over the total looking time to both objects. In some cases, researchers infer that more time looking at the target indicates better word comprehension (Sander-Montant, López Pérez, and Byers-Heinlein 2023; White and Morgan 2008), but often looking times significantly greater than chance (typically defined as looking to both images equally or similar to a baseline looking prior to word onset) are simply taken as evidence of comprehension, regardless of how much greater than chance they are (Swingley 2011).

The looking-while-listening paradigm has been found to be valid and reliable for measuring word comprehension when correlating global performance with aggregated vocabulary scores. For example, for both monolinguals and bilinguals, children with larger vocabulary sizes perform better on looking-while-listening measures (Fernald et al. 2006, 2013; Hurtado et al. 2014; F. Law and Edwards 2015; Marchman et al. 2020; Marchman and Fernald 2008; Peter et al. 2019; Zangl et al. 2005). For bilingual children, this relationship is largely language-specific. For instance, amongst Spanish-English bilingual toddlers, English vocabulary size is correlated with lexical processing in English looking-while-listening trials, while cross-language correlations tend to be weaker or absent (Hurtado et al. 2014; Marchman, Fernald, and Hurtado 2010). However, more recent research with Spanish-English bilinguals has identified some cross-language effects whereby better Spanish processing speed was related with later English language outcomes, suggesting that for bilinguals, good L1 processing may contribute to later L2 skills (Marchman et al. 2020).

1.1.3 | The Relative Strengths and Weaknesses of Parent Report and Behavioral Measures

Both parent report measures, like the CDI, and behavioral measures, like the looking-while-listening procedure, tap into children's word knowledge, but they excel in different areas (Fenson et al. 1993, 1994; Fernald et al. 2008). Parent reports like the CDI provide a relatively fast and inexpensive way to gather data about a large set of words from informants who are the primary source of children's language experience. Yet, there is evidence that caregivers can disagree about which words a child knows, because their reporting is inherently tied to their unique experiences with the child (De Houwer, Bornstein, and Leach 2005), and/or because of idiosyncratic inaccuracies in knowing, remembering, or reporting children's knowledge. In addition, the CDI can only provide a coarse estimate of knowledge for any given word, because response options are binary. It does not, for example, gather information on the regularity and range of situations in which a child uses the word (e.g., whether "water" has been produced only once at bath time, or whether it is regularly produced also when a child is thirsty or sees the ocean). The CDI makes up for restricted reporting options for each word by querying hundreds of words.

In turn, behavioral measures like the looking-while-listening procedure strike the opposite balance: researchers can gather moment-to-moment direct data about children's comprehension of tested words, and measures like proportion of looking time are, in theory, continuous. However, experiments rely on children's willingness to participate, and typically only test children's knowledge of a few words (Bergelson 2020). Moreover, an infant's looking behavior is assumed to index their word knowledge, but looking behavior is also likely affected by additional variables, both known (e.g., attention, speech processing) and unknown (Aslin 2007). Thus, while the CDI excels in the quantity of words it can query, looking-while-listening shines in the quality of the data collected regarding children's knowledge of each individual word tested.

Furthermore, both the CDI and the looking-while-listening task grapple with a linking problem: although researchers generally agree that these measures assess some aspect of word knowledge, it is unclear what level of knowledge they capture (Styles and Plunkett 2009). Parents may have some uncertainty as to whether their child has truly understood and/or meaningfully produced a word when reporting on the CDI (Houston-Price, Mather, and Sakkalou 2007; Frank et al. 2021). For example, children are known to overextend the meanings of certain words relative to adult speech: a child might produce the word "ball" to refer to all round objects, not just toys (Rescorla 1980). This word would likely be reported by some parents as "produced" on the CDI. Similarly, in addition to the cognitive processes involved—such as visuomotor planning, visual processing abilities, and the efficiency of word retrieval (Loi et al. 2017)—above-chance performance in the looking-while-listening task might reflect either fragile or robust word knowledge. For example, the semantic relatedness of distractors to target words (Bergelson and Aslin 2017) as well as the frequency of the target word (Kartushina and Mayor 2019; Potter and Lew-Williams 2023) affect performance. Indeed, it is likely that the level of knowledge being tapped in these measures changes with age and language experience (Bergelson 2020; Sander-Montant, López Pérez, and Byers-Heinlein 2023).

1.2 | Measuring Children's Word-Level Vocabulary Knowledge

Beyond measuring children's aggregate vocabulary knowledge, measuring the individual words a child knows can also provide valuable insight into children's word learning. First, words vary in their meanings, syntactic roles, phonological difficulty, and in many other ways which have been shown to matter for learning when examined in aggregate (Braginsky et al. 2019; Goodman, Dale, and Li 2008). If current methods enable researchers to accurately determine individual children's knowledge of specific words, these features can be more precisely investigated in regard to the roles they play in learning. Second, theoretical frameworks of vocabulary development emphasize the idea that acquiring a word involves a certain number of learning instances (Mollica and Piantadosi 2017), and tests of such theories benefit from accurately pinpointing when or to what degree a child has acquired a word. Third, parent report measures are already used at the word level, even though little research has assessed the reliability and validity of doing so (Houston-Price, Mather, and Sakkalou 2007). Lastly, clinicians could also benefit from the measurement of individual words in order to identify specific targets for intervention (Yoder, Warren, and Biggar 1997).

However, only a few studies have assessed the reliability of measurement of individual children's word-level vocabulary knowledge. Yoder, Warren, and Biggar (1997) examined the test-retest reliability of word-level and aggregate receptive vocabulary scores on the CDI, by examining consistency over a 2-week period for 17 children aged 18–33 months with atypical development. Aggregate total vocabulary scores showed robust reliability ($g = 0.93$), whereas word-level scores showed only moderate reliability ($k = 0.47$). Houston-Price, Mather, and Sakkalou (2007) collected CDI and looking time data in two experiments: one with 29 infants aged 18–22 months and another with 113 infants aged between 15 and 21 months. Children significantly increased their looks to the labeled object for both words reported as "known" on the CDI (e.g., mean increase of 7%) as well as those reported "unknown" (e.g., mean increase of 3.8%), suggesting that parents may have underestimated children's word knowledge (see also Hendrickson et al. 2015; Potter and Lew-Williams 2023, for similar findings). Other studies have found that word-level CDI comprehension data does not improve model fit, especially when the tested words are challenging for that age group (Kartushina and Mayor 2019; Moore and Bergelson 2022).

Although these results cast some doubt on strong word-level reliability, other studies offer supporting evidence for it. Styles and Plunkett (2009) conducted a conceptual replication of the Houston-Price, Mather, and Sakkalou (2007) study with 35 18-month-olds, this time using more challenging word pairs (e.g., words of the same semantic category like *cup-bowl*). Infants only showed a significant increase in target looks for words they were reported to understand (mean increase of 10%) and no increase in target looks for words they were reported to not understand, which was interpreted as evidence of word-level reliability and validity. Additional evidence for word-level reliability was reported in a touchscreen task, where bilingual children performed better on word comprehension trials where parents reported that they could produce the target word ($M = 65\%$) than on trials where parents reported that they could not produce the word ($M = 52\%$; Poulin-Dubois et al. 2013; see also moderate correlations reported between looking and pointing on single trials in Creel 2024). More recent preliminary reports have also found convergent validity between parent reports and looking-while-listening performance at the item level (e.g., Chai, McDonald, and Ko 2024; Smolík et al. 2023; Weaver and Saffran 2024).

1.3 | Child-Level Factors That Predict Children's Word Knowledge

In interpreting the mixed results of studies assessing the measurement of children's word-level knowledge, it is crucial to consider the potential drivers of individual differences in the specific set of words children know. First, word knowledge is driven by a child's idiosyncratic experiences, such as whether a child first produces "mommy," "daddy," or "grandma" depending on the caregiver they spend the most time with (Laing and Bergelson 2020), as well as the specific words they learn based on their own personal interests (e.g., a greater interest in vehicles than animals; Ackermann, Hepach, and Mani 2020). Second, word knowledge writ large is driven by other more universal child-level factors like age, total vocabulary size, and amount of language experience. Evidence that our measures are sufficiently sensitive to capture idiosyncratic word knowledge would come from our ability to detect variability in word knowledge above and beyond the effects of child-level factors.

Below, we will review evidence that age, total vocabulary size, and language experience are related to children's word knowledge, as reflected in their performance on the CDI and/or looking-while-listening task, thus motivating our consideration of these child-level factors as a potential explanation for item-level correlations between measures.

1.3.1 | Age

Children's capacity to learn new words and process known words improves with age, with parent reports indicating large increases in receptive and productive vocabulary size (Fenson et al. 1994), and faster, more accurate word recognition in behavioral tasks in older children compared to younger children (Fernald et al. 2006; Peter et al. 2019; Sander-Montant, López Pérez, and Byers-Heinlein 2023; although age effects in children's looking patterns are not always found, F. Law and Edwards 2015). Theories attribute this improvement to the gradual refinement of cognitive and linguistic processes like segmenting words from the speech stream, pairing them to referents, storing them in memory, and recalling them later (Samuelson 2021). Furthermore, enhanced cognitive, social, and linguistic skills develop with age. For example, as children age, they show gains in joint attention, an enhanced ability to track social partners' knowledge states, better phonological representations, and more efficient use of speech patterns to aid in sentence processing (Bergelson 2020; Meylan and Bergelson 2022). Thus, all other factors being equal, an older child should be more likely to know a particular word than a younger child.

1.3.2 | Language Experience

Language experience contributes importantly to building children's vocabularies and is typically measured in one of two ways: total quantity of speech and relative language exposure. Total quantity of speech is often measured via at-home language recordings, and expressed as the number of words a child hears per unit time (e.g., words per hour or per day) as counted from either hand transcription or automated algorithms (Cristia et al. 2021; Ganek and Erik-Brophy 2018). Research has shown that monolingual children who hear a greater quantity of child-directed speech at home have larger vocabularies and better in-lab word recognition (Weisleder and Fernald 2013), and bilinguals show faster word processing and perform better on standardized tests when they hear more words in a particular language (Marchman et al. 2017).

Relative language exposure is expressed in terms of a percentage of time that a child hears a particular language and is typically measured via detailed caregiver interviews or questionnaires, which are easier to obtain than at-home language recordings. Moreover, relative language exposure varies continuously from 0% to 100%, encompassing fully monolingual experience with a single language (100% exposure to one language, 0% to another), perfectly balanced bilingual experience (50% exposure to each language), and everything in between. Children with more relative exposure to a language tend to know more words in that language and show better word processing (Byers-Heinlein et al. 2024; Hurtado et al. 2014; Legacy et al. 2018; Pearson et al. 1997; Place and Hoff 2011), although there is some evidence that total quantity of speech is more predictive than relative exposure (Marchman et al. 2017). In sum, all factors being equal, a child with more language experience (whether assessed in terms of speech quantity or relative exposure) in a given language would be more likely to know a particular word in that language than a child with less experience.

1.3.3 | Vocabulary Size

Children's current vocabulary size is closely related to the specific words they are likely to know, and this connection may be due to several factors. First, children with larger vocabularies tend to process and recognize words more quickly and accurately than children with smaller vocabularies (Marchman, Fernald, and Hurtado 2010). Second, words are not learned randomly as there is semantic organization in a child's lexicon and vocabulary learning builds on itself: Children with larger vocabularies have denser and more interconnected semantic networks and they can use the words they know to facilitate the processing of new and known words (e.g., incorporating words that share perceptual features or semantic overlap more easily; Beckage, Smith, and Hills 2010; Borovsky et al. 2016; Borovsky 2022a, 2022b). Third, many words assessed in looking paradigms are from the CDI, a closed list of common words that children eventually learn almost entirely. Therefore, the more words a child is reported to know, the higher the likelihood that any randomly selected word from the CDI will also be known. For all of these reasons, when comparing two hypothetical children, both 24 months old—one with a vocabulary size in the ~10th percentile (70 words) and another in the ~90th percentile (500 words)—it would be reasonable to believe that a child with a larger vocabulary would be more likely to know a later-acquired word like "zipper," or any other particular word, than a child with a smaller vocabulary.

In sum, children's early vocabularies can be highly variable in size and content, with child-level factors like age, experience (whether in terms of total speech quantity or relative language exposure), and total vocabulary size each potentially predicting children's word knowledge, whether assessed via CDI or looking-while listening.

1.4 | The Current Study

This study aimed to assess the reliability and validity of current methods in measuring individual infants' word-level knowledge. For measures like CDI and looking-while-listening to be reliable and valid at the item level, we would expect that first, parents must be accurate reporters of their child's individual word knowledge. Second, looking-while-listening must be a reliable and valid way to assess infants' idiosyncratic knowledge of individual words. If both conditions are sufficiently met, then parent reports and behavioral measures of word-level knowledge should be related. If the relationship remains even when we control for child-level factors like age, language exposure, and vocabulary size, then this would suggest that we are capturing word-specific variance in knowledge above and beyond what would be expected for any child with a similar linguistic and developmental profile. However, if parent-reported word knowledge does not explain any additional variance in looking-while-listening performance beyond child-level factors, then this would suggest that, at the item level, these measures capture something general about development patterns but not something specific about individual children's vocabularies. We examined these competing predictions in French-English monolingual and bilingual children: (1) in our full sample of 126 children aged 14–31 months, for whom we had CDI word production and looking-while-listening data available and (2) in a small subsample of 30 children, all approximately 14 months old, for whom we had CDI word comprehension and looking-while-listening data available.[1] Our sample includes variability across all three of the child-level factors we examined: Participants varied in their ages, their reported vocabulary sizes, and their percent exposure to the tested language. This variability in the sample helped us disentangle the relationships between these variables and word knowledge.

2 | Method

The present study was conducted according to guidelines laid down in the Declaration of Helsinki, with written informed consent obtained from a parent or guardian for each child before any assessment or data collection. Ethics approval was granted by the Human Research Ethics Committee of Concordia University (certification #10000439). The current study was preregistered[2] via the Open Science Framework at https://osf.io/ygsnf. The analyzed dataset is a subset of the archival data used in Sander-Montant et al. (2023; https://osf.io/2m345/), retaining three studies (out of a possible five) where both CDI parent report data and looking-while-listening data were collected. The studies were conducted at the Concordia Infant Research Lab in Montreal, Canada, between 2012 and 2019 (Byers-Heinlein, Morin-Lessard, and Lew-Williams 2017; Schott, Moore, and Byers-Heinlein 2022; unpublished). Each child was tested in one of their native language(s), which in one study was English for all participants (Schott, Moore, and Byers-Heinlein 2022), and in two studies was English or French (randomly assigned; Byers-Heinlein, Morin-Lessard, and Lew-Williams 2017; unpublished data). Data, analysis code, and appendices are available on the Open Science Framework at https://osf.io/mxksz/.

2.1 | Participants

The final sample consisted of 127 testing sessions from 126 participants (72 males, 54 females). One child contributed both CDI and eyetracking data at two ages (14 and 20 months). In the subsample with parent-reported comprehension data, children were much younger ($M = 14$ months, $SD = 12$ days) with a tight age range (13.8–15.5 months). In the full sample with parent-reported production data, children had a wider age range, from 14 to 31 months ($M = 22$ months, $SD = 5$ months).

All children were living in the Montreal area and were acquiring French and/or English (details on language exposure are provided in the next section). A demographics questionnaire gathered data on child ethnicities, where parents could choose from a range of categories or indicate another ethnicity. The most commonly reported ethnicity was European (48%), followed by unknown or unspecified ethnicities (16%). The next category involved indicating multiple ethnicities (15%; e.g., Arab-European, African-European, and Latin American-Canadian, etc.). This was followed by "other" (13%), for instance, Canadian. Caregivers had high levels of education: 62% of mothers and 58% of fathers had obtained a bachelor's degree or higher.

No further participant-level exclusions were made beyond those made by Sander-Montant et al. (2023) given that this dataset had already excluded participants due to premature birth (< 37 gestational weeks), low birth weight (< 2500 g), health issues that could affect performance (e.g., ear infection at time of testing), a developmental delay diagnosis, being exposed to a third language more than 10% of the time, and children who had been exposed to their second language for less than half their life.

2.2 | Measures

2.2.1 | Parent-Reported Vocabulary Measure

Children's vocabulary in French and/or English (depending on the language(s) the child was acquiring) was measured using the CDI (Fenson et al. 1993) and its Quebec-French adaptation (Trudeau, Frank, and Poulin-Dubois 1999). For children under 18 months, parents indicated whether their child understood and/or produced each word listed using the Words and Gestures form ($n = 30$ children, all around 14 months). Meanwhile, for children 18 months and older, parents only indicated whether their child produced each word listed using the Words and Sentences form ($n = 97$ children; one child participated in two studies and therefore contributed data from both forms). These assessments provided word knowledge scores for each word on the form (understands/produces word or does not understand/produce word) as well as measures of total vocabulary size for children's respective language(s).

Parent-reported comprehension is conceptually closer to looking-while-listening performance (which also measures comprehension), but as this was only assessed for the subset of children who completed the Words and Gestures form, analysis of this variable was exploratory. As both forms assess production, this measure was available for all participants and thus was analyzed as our primary parent-report vocabulary measure.

Parent-reported word comprehension and production were operationalized as children's CDI-reported score for each tested word, with separate scores for comprehension (0 = not understood, 1 = understood) and production (0 = not produced, 1 = produced). For the subset of children with parent-reported comprehension data, children were reported on the CDI to understand, overall, 52% (SD = 30%, range = 0%–100%) of the tested target words in the looking-while-listening task (149/290 trials). In the full sample, on average, children were reported on the CDI to produce 45% (SD = 39%; range = 0%–100%) of the tested target words in the looking-while-listening task (642/1364 trials).

Total vocabulary size was operationalized as children's total receptive/expressive vocabulary size, excluding the words they were tested on in the looking-while-listening task to avoid redundancy from word-level information, thus preventing any overlapping variance. In the comprehension subsample, receptive vocabulary ranged from 0 to 280 ($M = 96$, $SD = 61$). For the full sample, expressive vocabulary ranged from 0 to 648 ($M = 159$, $SD = 174$).

2.2.2 | Language Exposure Questionnaire Using MAPLE

Participants' language experience was measured in terms of relative language exposure, using the Language Exposure Questionnaire (Bosch and Sebastián-Gallés 2001) with the Multilingual Approach to Parent Language Estimates (MAPLE; Byers-Heinlein et al. 2020). This structured interview captures a child's lifetime exposure to each language as a percentage. Children's exposure to the tested language was analyzed on a continuum, as this provides nuance beyond traditional monolingual/bilingual dichotomies (e.g., Kremin and Byers-Heinlein 2021; Rocha-Hidalgo and Barr 2023). Exposure to the tested language in the comprehension subsample ranged from 24% to 91% ($M = 51\%$, $SD = 17\%$) and in the full sample ranged from 12% to 100% ($M = 63\%$, $SD = 25\%$).

2.2.3 | Looking-While-Listening Measure

Looking-while-listening data were gathered using a Tobii T60-XL eyetracker with Tobii Studio software. During the task, parents wore darkened sunglasses and/or listened to music on headphones, which helps prevent parental awareness from impacting infants' behavior (Alcock, Watts, and Horst 2020). On each trial, children saw yoked pairs of target-distractor images (e.g., dog-book; target counterbalanced; for a complete stimulus list see Appendix B), and heard the target labeled (e.g., "Look! Find the dog!"). All target words were from the CDI, selected because they were likely to be understood at the tested ages (e.g., Frank et al. 2017). Trials with manipulations like mispronunciations were excluded, and only control or filler trials where words were presented in a typical carrier sentence were retained (e.g., "Look! Find the ___!"; 12 possible trials per infant), yielding 1364 analyzed trials.

3 | Results

3.1 | Dependent Variable: Proportion Looking to Target

Following Sander-Montant et al. (2023), we delineated two broad AOIs (left and right) that together encompassed the entire screen—a method previously used with good outcomes (Hessels et al. 2016). The dependent variable was participants' looking directed toward the target AOI within the analysis window of 360–3000 ms after the target noun onset, divided by their total looking time to either AOI. We designated a longer window of analysis than the archival studies analyzed (typically 360–2000 ms) given that longer analysis windows (beyond 2000 ms post-target onset) have been found to be more reliable for detecting individual differences (Zettersten et al. 2021). No further trial-level exclusions were made beyond those previously made by Sander-Montant et al. (2023), which had already excluded trials involving experimental manipulations, eyetracker malfunction (e.g., failed to capture participants' gaze), testing errors (e.g., interference by researcher or parent), instances with invalid gaze coordinates, and trials with low validity codes automatically generated by the eyetracker.

In the subsample of children with comprehension data, on average, children looked to the labeled target 53% ($SD = 11\%$, range = 28%–75%) of the time. In the full sample with production data, on average, children looked to the labeled target 63% ($SD = 12\%$, range = 28%–100%) of the time. Additional details about children's distribution across predictor and outcome variables can be found in Figure 1.

3.2 | Correlation Analysis

Given that previous studies (e.g., David and Wei 2008; Hurtado, Marchman, and Fernald 2007) have found associations between our key variables (IVs: target word comprehension/production [binary], age [in months], language exposure [in percentage], and total vocabulary size; DV: proportion target looking), prior to delving into the main analyses, we conducted a correlation analysis to assess their relationships (see Figure 2). A Shapiro-Wilk test suggested non-normality, therefore, we conducted non-parametric Spearman's rank correlations, although Pearson's correlations yielded similar results.

Comprehension: Child-Level Correlations in Subsample. At the child level, results showed that there was one statistically significant correlation amongst predictors after adjusting for multiple comparisons: children who were reported to comprehend more of the tested words had larger total vocabulary sizes ($r = 0.86$). Notably, no significant correlations between predictors and proportion of target looking were found, thus we do not report further analysis of the comprehension data at the item level.

Production: Child-Level Correlations in Full Sample. At the child level, there were statistically significant moderate-to-high correlations amongst all predictors and with the outcome variable, even after adjusting for multiple comparisons. For example, children who produced more of the tested words were older ($r = 0.67$), had more exposure to the tested language ($r = 0.46$), and produced more non-tested CDI words ($r = 0.91$). Furthermore, all predictors had moderate correlations with our outcome variable, proportion target looking ($rs = 0.37$–0.60).

3.3 | Model Construction at the Trial Level

To test the robustness of word-level production data beyond age, language exposure, and total productive vocabulary size, we ran a series of pre-registered linear mixed-effects models using the *lme4* package in R (Bates et al. 2015) with random effects for subjects

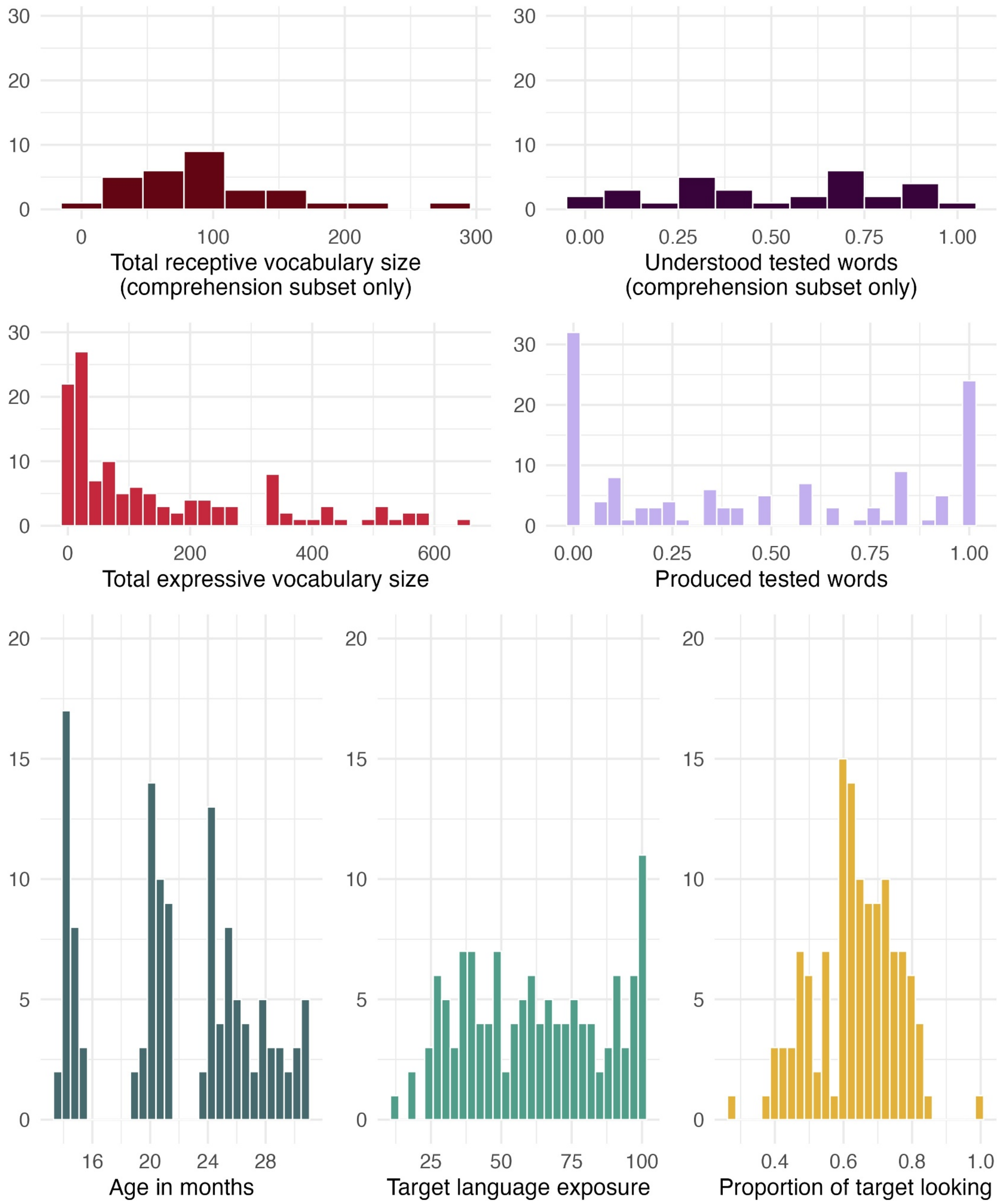

FIGURE 1 | Histograms of predictor and outcome variables. Histograms illustrate the distribution of our predictor and dependent variables across the subsample and the full sample, with counts on the *y*-axis. For the subsample of 14-month-olds with comprehension data, their total receptive vocabulary size and the proportion of tested words reported to be understood (labeled "understood tested words") on the CDI are shown on the top. All other figures correspond to the full production sample which includes all 126 children tested. Their displayed data encompasses children's total expressive vocabulary size, the proportion of tested words that children were reported to produce on the CDI (labeled "produced tested words"), children's age in months, children's language exposure to the tested language, and children's mean proportion of target looking on the looking-while-listening task.

and target items. Model comparisons were then conducted using the *anova()* function from the *stats* package in R (R Core Team 2021) to find the best-fitting model. This approach began with running simple models in which produces_word (whether the child was reported to produce the target word) and vocabulary_size (total CDI vocabulary score) were the respective sole predictors of children's trial-level target looking. We then increased model complexity by creating nested models that accounted for the possibility of joint effects of *produces_word* + *vocabulary_size* as well as a *produces_word*vocabulary_size* interaction. Subsequently, we created further nested models where we added age and language_exposure (percent of exposure to the target language) as fixed effects (for all model syntax, see Appendix C). Model comparisons were conducted to find the best-fitting model. In the next section, we detail the outcomes of the produces_word-only model, followed by the model comparisons. We report both standardized effect sizes (β) which range from -1 to $+1$ and can be straightforwardly compared to the Spearman correlation coefficients presented above, as well as unstandardized effect sizes (b) which are in the familiar metric of proportion looking to target (e.g., a predictor with $b = 0.03$ is associated with 3% more target looking).

3.3.1 | Word Production

Word Production-Only Model Outcome. There was a positive and statistically significant effect of produces_word on looking-while-listening performance (see Table 1), where being reported to produce the tested word was associated with a 9% increase in children's proportion of target looking (see Figure 3).

Word Production Model Comparisons. Next, we conducted model comparisons to assess whether in the best-fit model, produces_word was still a significant predictor when other predictors were added. Model comparisons were performed via likelihood-ratio chi-squared tests, and fit was further assessed using R-squared, the Akaike Information Criterion (AIC; Akaike 1974), and the Bayes Information Criterion (BIC; Schwarz 1978; see Appendix C for model outputs). The best-fitting model included produces_word, age, and vocabulary_size as predictors of children's performance, and did not include language_exposure. The produces_word + age + vocabulary_size model fit significantly better than the model with only vocabulary_size (X^2 $(2) = 15.944$, $p < 0.001$), the model with only produces_word (X^2 $(2) = 23.896$, $p < 0.001$) and the model with produces_word + vocabulary_size (X^2 $(1) = 12.01, p < 0.001$). Though not significantly different, this model showed lower AIC (558.76) and BIC (595.29) values than more complex models containing interactions (i.e., produces_word*vocabulary_size and produces_word*vocabulary size + age/language_exposure), indicating better fit. Thus, we denote this as our best-fitting model.

Word Production Best-Fitting Model. Examination of coefficients in the best-fitting model (Table 2) suggest that, when other predictors were held constant, being reported to produce a tested item predicted a non-significant 3% increase in target looks (compared to the 9% increase in the produces_word-only model). Standardized effect size estimates suggest that age had the largest effect on looking-while-listening performance, followed by produces_word, and then vocabulary_size (see Table 2). The standard error for produces_word was much larger than for the other two predictors, suggesting that produces_word was a more variable predictor. The fixed effects of produces_word, age, and vocabulary_size, cumulatively explained 6% of the variance in children's performance on this task and together with the random effects accounted for 17% of the variance in their performance (see Figure 4). Notably, the magnitude of the effects in this model were similar and the model was not significantly different to that of an age + vocabulary_size model which omits produces_word (see Appendix D for model effects).

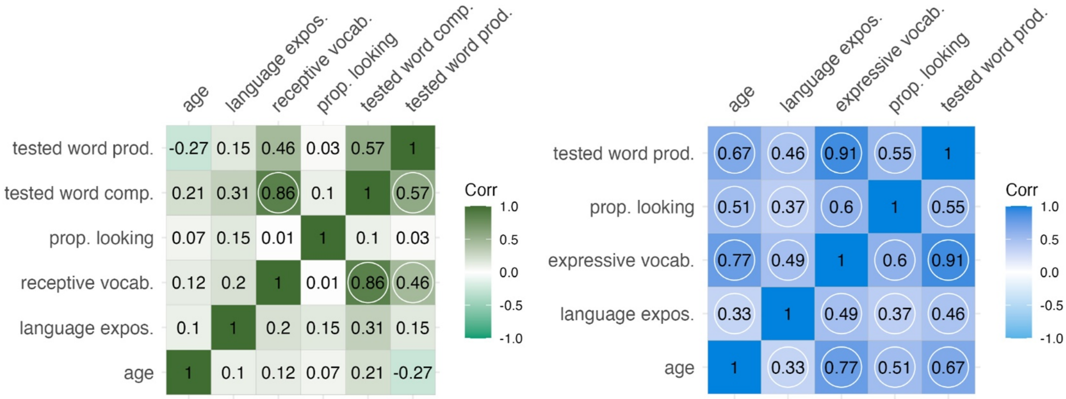

FIGURE 2 | Child-level correlation plot of predictor and outcome variables. Child-level Spearman correlation matrix for the subsample with comprehension data (left) and the full sample with production data (right). For the comprehension subsample, all statistically significant correlation coefficients are circled in the lower triangular part of the correlation matrix. Correlations that are still significant after correcting for multiple comparisons are circled in the upper triangular part of the correlation matrix. For production (right), all correlations are circled because they were statistically significant both before and after correcting for multiple comparisons. Darker shading indicates stronger correlation. "Tested word comp." and "tested word prod." refer to on average, how many of the **tested** words children were reported to understand or produce on the CDI. "Prop. looking" refers to children's average proportion of target looks on the looking-while-listening task. "Receptive vocab." and "Expressive vocab." refer to children's total receptive or productive vocabulary size as captured on the CDI, excluding the tested items. "Language expos." refers to children's exposure to the tested language as a percentage of total language exposure. "Age" refers to children's chronological age (in days for comprehension given the tight age range and in months for production).

TABLE 1 | Fixed effect output of a produces_word model.

Fixed effects	b	β	SE	df	t	p	
Intercept	0.58	−0.16	0.021	1359	28.02	< 0.001	Cond R^2 = 0.13 Marg R^2 = 0.02
produces_word	0.09	0.29	0.018	1359	4.93	< 0.001	

Note: This table presents the results of the **produces_word** model. The displayed coefficients include unstandardized betas (b), standardized betas (β), standard errors (SE), degrees of freedom (df), test statistic (t), and p-values (p). Additionally, the conditional (Cond) and marginal (Marg) R^2 values are shown.

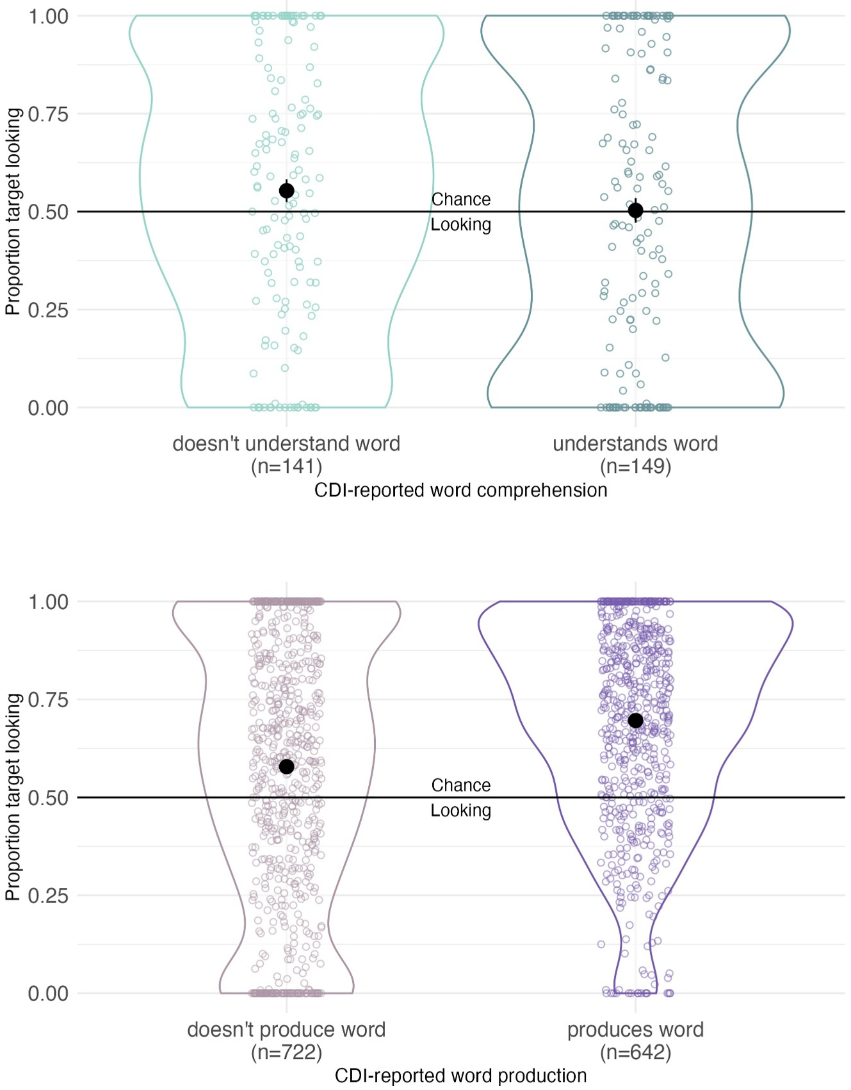

FIGURE 3 | Raw values of children's looking accuracy by reported word knowledge. Raw values of children's proportion of target looks on the looking-while-listening task grouped by CDI-reported word comprehension (top) and production (bottom). Individual points represent one participant's performance on an individual trial. Black pointranges show the group means with standard error (note that for production, error bars are small and subsumed within the black points). The reference line at 0.5 indicates chance looking.

TABLE 2 | Fixed effect output of a produces_word + age + vocabulary_size model.

Fixed effects	*b*	*β*	SE	df	*t*	*p*	
Intercept	0.39	−0.04	0.055	171.0	7.157	< 0.001	Cond R^2 = 0.17 Marg R^2 = 0.06
produces_word	0.033	0.11	0.023	1038	1.452	0.147	
age	0.0091	0.15	0.0026	186.6	3.481	< 0.001	
vocabulary_size	0.00015	0.08	0.000075	174.6	2.035	0.043	

Note: This table presents the results of the best-fitting model, namely the **produces_word** + **age** + **vocabulary_size** model. The displayed coefficients include unstandardized betas (*b*), standardized betas (*β*), standard errors (SE), degrees of freedom (df), test statistic (*t*), and *p*-values (*p*). Additionally, the conditional (Cond) and marginal (Marg) R^2 values are shown.

4 | General Discussion

This study examined the measurement of infants' knowledge of individual words via parent reports (CDI) and behavioral measures (looking-while-listening). We analyzed data from 126 French and/or English monolingual and bilingual 14–31-month-olds with parent-reported production data, and a subsample of thirty 14-month-olds with parent-reported comprehension data. We assessed whether children who were reported to understand and/or say a given word (e.g., "ball")

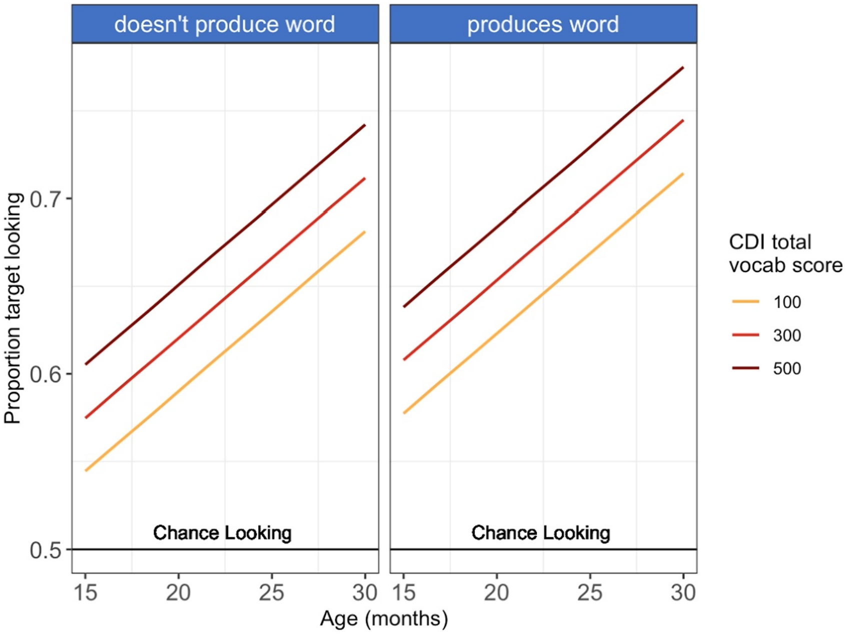

FIGURE 4 | Model-predicted values of looking accuracy (proportion target looking) by age and total vocabulary, split by whether or not the child was reported to produce the target word. Predicted values of proportion target looks (*y*-axis) from the best-fitting model are shown, which included **produces_word** + **age** + **vocabulary_size** as predictors. On the *x*-axis is children's age in months. The colored lines indicate total productive vocabulary size bins depicting a range from 100 to 500 words. Figure is faceted by reported production of the target word, although this predictor was not significant in the model.

spent more time gazing at that object when it was a labeled target (e.g., "Look at the ball!") than children who were not reported to understand and/or say the word. Parent-reported word production alone significantly predicted trial-level looking-while-listening performance in the full sample ($\beta = 0.29$), although we did not find the same relationship with parent-reported comprehension in the subsample. Nonetheless, the relationship for production could suggest convergent validity between the two methods, as has been the conclusion in previous work (Poulin-Dubois et al. 2013; Styles and Plunkett 2009; see also preliminary reports from Chai, McDonald, and Ko 2024; Smolík et al. 2023; Weaver and Saffran 2024).

We extended beyond previous work to control for three child-level variables that could also explain this association: children's age, language exposure, and total vocabulary size. Indeed, in the winning predictive model, age ($\beta = 0.15$) and total productive vocabulary ($\beta = 0.08$) were significant predictors of trial-level looking-while-listening performance. Yet, once they were controlled for, parent-reported target word production had a much smaller effect than when it was the sole predictor (reduction from $\beta = 0.29$ to $\beta = 0.11$), and was no longer statistically significant. Language exposure did not improve model fit, suggesting that any effects of language exposure were captured through other predictors. In general, fixed effects only captured 6% of the variance in target looking. Together, our findings suggest that current approaches to measuring children's knowledge of individual words—whether via looking time or parent report—may lack sufficient reliability to assess children's idiosyncratic knowledge. In the next sections, we explore the interconnected issues that may explain this pattern of results.

4.1 | On Measurement: Reliability in Parent Report and Looking Time Measures

We first consider the pairwise relationship between parent-reported vocabulary production and looking-while-listening performance. Ignoring other predictors, the proportion of time that children looked at a labeled target was linked to whether they were reported to say that particular word, but the effect was weaker when data were analyzed at the trial level ($\beta = 0.29$) than when data were averaged for each child then correlated ($r = 0.55$). While item-level correlations have generally not been reported in the literature (instead studies have reported average performance for parent-reported known vs. unknown words, e.g., Houston-Price, Mather, and Sakkalou 2007; Styles and Plunkett 2009), studies examining correlations between looking-while-listening performance and total vocabulary[3] report similar magnitude of effects as in our study ($r = -0.20$ to -38 for groups of monolinguals aged 15–25 months, Fernald et al. 2006; $r = -0.41$ to -0.63 in 30-month-old bilinguals, Marchman, Fernald, and Hurtado 2010; -0.29 to -0.45 in 15–31-month-old monolinguals, Peter et al. 2019).

At first blush, this pattern appears to show a puzzling discrepancy between seemingly reliable aggregate measures of word knowledge and seemingly less reliable word level measures. Yet, this pattern of results exemplifies a well-established phenomenon in psychometrics, whereby aggregating a large number of measurements results in a more reliable measurement compared to a single item (Raykov and Marcoulides 2011; Spearman 1904). This phenomenon is evident when comparing the reliability of single-item measures to aggregated measures. For example, whether a child produces a single item, or whether a child looks to a target in 1–3 trials is a comparatively weaker measure than aggregated measures, either within an individual by summing across words to calculate a total vocabulary size, or within a single item by averaging across infants to determine the average age of acquisition of a particular word.

This is because any measurement is a combination of an individual's true ability score and some degree of measurement error (Spearman 1904). An aggregate score across many items will have less measurement error than any single item, thus better detecting an individual's true ability on the assessed construct, resulting in a more reliable measurement.

In the context of parent-reported vocabulary, children's reported knowledge of a specific word, as measured by a single binary data point, is less psychometrically reliable than a total vocabulary size score which averages across hundreds of individual responses (as seen in Yoder, Warren, and Biggar 1997). Summing or averaging across multiple items offsets inaccuracies like mistakenly marking that a child produces a word when they do not or marking that a child does not understand a word they do (for additional discussion, see Houston-Price, Mather, and Sakkalou 2007). Our own results are consistent with this pattern: our data suggested that individual word production had a standard error more than 300 times larger than the standard error for productive vocabulary size, underscoring that total vocabulary size offers a more reliable measurement than the measurement of any specific word. Indeed, averaging across just a handful of items provides a more reliable measurement than a single item: we found a child-level correlation of 0.91 between parent-reported average production of items tested in the looking-while-listening procedure and the rest of the items queried on the CDI. This same logic also applies to the looking-while-listening task, where aggregating across more items will yield a more reliable measurement than performance on one or two items. Here, measurement error is introduced by factors like the visual salience of images displayed on screen, side bias, familiarity with distractor items, preferences for one object over the other, tiredness during the task, and eyetracking issues (Aslin 2007; DeBolt and Oakes 2022).

The issue of obtaining reliable measurements plagues infant research as a whole (Byers-Heinlein, Bergmann, and Savalei 2022; for an example see Schreiner et al. 2024), but there may nonetheless be several avenues for decreasing measurement error to boost the reliability of measurement. For parent-reported word knowledge, one approach would be to gather word-level data that averages across more measurements, for example, by asking multiple informants to report children's word knowledge (e.g., De Houwer, Bornstein, and Leach 2005). A second approach could be to use continuous rather than binary scoring for parent-reported word knowledge (e.g., a Likert scale where for each word, parents rate how certain they are that their child understands or says each word), which could mitigate the reliability issues of binary items by offering parents a broader scale on which to place estimates of their children's word knowledge. Finally, for looking-while-listening, assessing the same item across more trials has been suggested as a way to improve reliability (e.g., multiple trials testing "dog"; Byers-Heinlein, Bergmann, and Savalei 2022), although there may be some limits in how many trials infants can participate in successfully.

4.2 | What Predicts Performance in Online Word Comprehension?

Moving beyond measurement challenges, we next delve into the factors that predicted infants' online word comprehension in our study. Child-level factors, namely age and total vocabulary size, were more robust predictors of looking-while-listening performance than children's parent-reported knowledge of the word tested on each trial. We offer several possible explanations for this finding. One possibility is that both age and total vocabulary might independently be similar or even better proxies of word knowledge than word-level parent reports. For example, "dog" is on average first produced at 16 months (when infants have a mean vocabulary size of < 25 words), "ear" at 18 months (when infants have a mean vocabulary of around 70 words), and "table" at 23 months (when infants have a mean vocabulary of around 300 words; Fenson et al. 1994; Frank et al. 2017). Therefore, including age in our model may have captured these age-of-acquisition effects better than parent reports of infants' individual word knowledge. Adding vocabulary size to the model may capture the vocabulary level at which a word is likely to be acquired. In this way, these child-level factors may provide a reasonable estimate of whether a child is likely to know a given word, perhaps because idiosyncratic differences in individual children's vocabulary knowledge are relatively minor.

Another factor that contributes to our findings is substantial shared variance between our predictors (*r*s between child-level word production, age, relative language exposure, and total vocabulary size ranging from 0.33 to 0.91), which all showed moderate-to-strong correlations with looking-while-listening performance (*r*s = 0.37–0.60). This finding is in line with previous research that has shown similarly strong correlations between age and productive vocabulary size ($r = 0.71$ in Zangl et al. 2005; $r = 0.82$ in Hurtado, Marchman, and Fernald 2007), language exposure and productive vocabulary size ($r = 0.65$ in David and Wei 2008; $r = \sim 0.70$ in Legacy et al. 2018), and language exposure and word recognition tasks ($r = -0.57$ in Hurtado et al. 2014). Shared variance can explain why the effect of word production on looking-while-listening performance was weak or absent in models including total vocabulary size, and why language exposure was not predictive when vocabulary size was considered. Overall, our findings highlight challenges in disentangling the distinct independent effects of multiple related factors that predict the words children know.

A complementary possibility is that age and total vocabulary size index children's general language abilities in a way that is predictive of their looking-while-listening performance. Indeed, age has been linked to vocabulary size (e.g., Zangl et al. 2005), as well as looking-while-listening performance (Fernald et al. 2006; Peter et al. 2019; Sander-Montant, López Pérez, and Byers-Heinlein 2023). Similarly, vocabulary size is predictive of a range of language abilities in early childhood, including children's ability to learn similar-sounding words (Werker et al. 2002), understand muffled speech (Zangl et al. 2005), and learn new words (F. Law and Edwards 2015). Further bolstering this second possibility and aligning with our general findings, research using longitudinal CDI word data from 15 to 36-month-olds found that models with age and total vocabulary size (and, in contrast to our findings, reported word production) significantly outperformed age-of-acquisition norm models in predicting children's future lexical knowledge (Beckage, Mozer, and Colunga 2015; although see F. Law and Edwards 2015, for an effect of vocabulary size but not age on aggregate scores). The authors suggested that both child-level and word-level factors capture overlapping but distinct information that can predict children's individual word knowledge beyond what age of acquisition norms capture.

In light of these cumulative findings, it becomes plausible that beyond mere word knowledge, age and total vocabulary size could reflect children's general language competence. Children with large vocabularies and older children tend to be more experienced word learners, with more developed cognitive skills that allow for more robust representation of words and their meanings. This raises the question of whether behavioral methods like the looking-while-listening task more closely index specific word knowledge or general language ability. Experimental manipulations of children's word knowledge, for example, comparing children's word-level performance on familiar and newly-taught words, might provide some leverage on this issue.

4.3 | Strengths, Limitations, and Future Directions

This study contributes to the limited body of literature assessing whether current methods used to capture children's vocabulary knowledge can accurately measure children's knowledge of individual words. A strength of our study is the use of data from a large sample with a wide age range and with a wide range of language exposure, including both monolinguals and bilinguals. This resulted in greater variation in infants' knowledge of individual words as well as high statistical power, which improved our ability to disentangle different predictors of performance. At the same time, the inclusion of bilingual participants could potentially increase measurement error as parents had to assess their children's word knowledge in two languages. However, this concern is somewhat attenuated by results from previous research, as well as our own findings, suggesting that the CDI has good reliability and validity for bilinguals (Marchman and Martínez-Sussmann 2002), at least at the aggregate level.

It is also important to note that while the looking-while-listening task measures word comprehension, our primary measure of parent-reported word knowledge was based on word production. Parent-reported comprehension was only available for a subset of participants, all of whom were around 14 months of age, and this smaller sample with a more restricted range could explain why we did not find a relationship between this variable and looking-while-listening performance. Further, while both parent-reported comprehension and production each provide a measure of word knowledge, it is unclear the degree to which they tap into distinct facets of lexical knowledge. Previous work has shown that comprehension and production vocabulary size scores are moderately to strongly correlated (e.g., $r = 0.34$–0.67; Feldman et al. 2000; Marchman et al. 2018), and this was also the case in our subsample dataset ($r = 0.46$, $p = 0.001$). However, differing perspectives exist; while some argue that comprehension and production are clearly distinct (Benedict 1979; Keenan and MacWhinney 1987), others propose that they are unitary or interwoven (e.g., Chater, McCauley, and Christiansen 2016; Pickering and Garrod 2013). To further examine the relationship between comprehension and production, future investigations should measure both parent-reported word comprehension and word production at a wider range of ages to see if and how results vary.

Another limitation is that we were only able to analyze the specific items tested in the looking-while-listening task within our archival dataset, which were chosen to be as easy as possible given experimental constraints. Although children in our sample were not usually at ceiling in their performance, the inclusion of more difficult items could produce greater variability, which in turn might reveal stronger relationships. Relatedly, because our study used archival data, different infants were tested on different items, although potential differences across items were statistically controlled for using random effects. Future studies could more systematically manipulate item difficulty (see Fibla Reixachs 2021), either by testing infants of different ages on an identical set of items with different difficulties, or by setting a standard set of difficulty levels and using different items at different ages.

5 | Conclusion

The results of our study indicated that parent-reported word production only moderately predicted experimentally-measured word comprehension at the word level, and performance was just as well or better predicted by child-level factors, specifically a child's age and their vocabulary size. This finding serves as a cautionary note for researchers intending to assess children's idiosyncratic knowledge of individual words as current methods may lack sufficient reliability, particularly when using single-item binary scores. Concurrently, the study sheds light on the interconnected nature of various child-level factors that play a role in children's word acquisition and the challenges in disentangling their independent effects. These issues signal a need for the development of new approaches to measure word knowledge in more reliable ways. Better methods could enable researchers to better link children's actual word knowledge to theories on word learning to further understand vocabulary development, while allowing clinicians to identify specific words for targeted intervention.

Author Contributions

Melanie López Pérez: conceptualization, methodology, formal analysis, writing–original draft, review and editing, visualization. **Charlotte Moore**: methodology, formal analysis, writing–original draft, review and editing, visualization. **Andrea Sander-Montant**: writing–review and editing. **Krista Byers-Heinlein**: conceptualization, methodology, writing–review and editing, supervision, funding acquisition.

Acknowledgments

We are thankful to all the families that contributed to this work. We thank the members of the Concordia Infant Research Lab and Dr. Ana Maria Gonzalez-Barrero for their support with this project. This project was funded by the Natural Sciences and Engineering Research Council (NSERC) to KBH (402470-2011; 2018-04390), as well as by a Concordia Faculty of Arts & Science Fellowship to MLP and by the Fonds de recherche du Québec—Société et culture to MLP. The authors declare no conflicts of interest with regard to the funding source for this study.

Data Availability Statement

The data, analysis code, and appendices that support the findings of this study are openly available on the Open Science Framework at https://osf.io/mxksz/.

Endnotes

[1] Correlations with comprehension data were conducted as an additional exploratory analysis.

[2] Refer to Appendix A for a detailed description of deviations from the pre-registration. All deviations were minor.

[3] Most studies comparing these two measures have done so via reaction time rather than proportion target looking, which is expected to have a negative correlation with vocabulary size.

References

Ackermann, L., R. Hepach, and N. Mani. 2020. "Children Learn Words Easier When They Are Interested in the Category to Which the Word Belongs." *Developmental Science* 23, no. 3: e12915. https://doi.org/10.1111/desc.12915.

Akaike, H. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19, no. 6: 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Alcock, K., S. Watts, and J. Horst. 2020. "What Am I Supposed to Be Looking at? Controls and Measures in Inter-Modal Preferential Looking." *Infant Behavior and Development* 60: 101449. https://doi.org/10.1016/j.infbeh.2020.101449.

Aslin, R. N. 2007. "What's in a Look?" *Developmental Science* 10, no. 1: 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67, no. 1: 1–48. https://doi.org/10.18637/jss.v067.i01.

Beckage, N., M. Mozer, and E. Colunga. 2015. "Predicting a Child's Trajectory of Lexical Acquisition." In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 190–195.

Beckage, N., L. Smith, and T. Hills. 2010. "Semantic Network Connectivity Is Related to Vocabulary Growth Rate in Children." In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, edited by D. C. Noelle, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, and P. P. Maglio, 2769–2774.

Benedict, H. 1979. "Early Lexical Development: Comprehension and Production." *Journal of Child Language* 6, no. 2: 183–200. https://doi.org/10.1017/S0305000900002245.

Bergelson, E. 2020. "The Comprehension Boost in Early Word Learning: Older Infants Are Better Learners." *Child Development Perspectives* 14, no. 3: 142–149. https://doi.org/10.1111/cdep.12373.

Bergelson, E., and R. N. Aslin. 2017. "Nature and Origins of the Lexicon in 6-Mo-Olds." *Proceedings of the National Academy of Sciences* 114, no. 49: 12916–12921. https://doi.org/10.1073/pnas.1712966114.

Borovsky, A. 2022a. "Drivers of Lexical Processing and Implications for Early Learning." *Annual Review of Developmental Psychology* 4, no. 1: 21–40. https://doi.org/10.1146/annurev-devpsych-120920-042902.

Borovsky, A. 2022b. "Lexico-Semantic Structure in Vocabulary and its Links to Lexical Processing in Toddlerhood and Language Outcomes at Age Three." *Developmental Psychology* 58, no. 4: 607–630. https://doi.org/10.1037/dev0001291.

Borovsky, A., E. M. Ellis, J. L. Evans, and J. L. Elman. 2016. "Lexical Leverage: Category Knowledge Boosts Real-time Novel Word Recognition in 2-year-olds." *Developmental Science* 19, no. 6: 918–932. https://doi.org/10.1111/desc.12343.

Bosch, L., and N. Sebastián-Gallés. 2001. "Evidence of Early Language Discrimination Abilities in Infants From Bilingual Environments." *Infancy* 2, no. 1: 29–49. https://doi.org/10.1207/S15327078IN0201_3.

Braginsky, M., D. Yurovsky, V. A. Marchman, and M. C. Frank. 2019. "Consistency and Variability in Children's Word Learning Across Languages." *Open Mind* 3: 52–67. https://doi.org/10.1162/opmi_a_00026.

Byers-Heinlein, K., C. Bergmann, and V. Savalei. 2022. "Six Solutions for More Reliable Infant Research." *Infant and Child Development* 31, no. 5: e2296. https://doi.org/10.1002/icd.2296.

Byers-Heinlein, K., A. M. Gonzalez-Barrero, E. Schott, and H. Killam. 2024. "Sometimes Larger, Sometimes Smaller: Measuring Vocabulary in Monolingual and Bilingual Infants and Toddlers." *First Language* 44, no. 1: 74–95. https://doi.org/10.1177/01427237231204167.

Byers-Heinlein, K., E. Morin-Lessard, and C. Lew-Williams. 2017. "Bilingual Infants Control Their Languages as They Listen." *Proceedings of the National Academy of Sciences* 114, no. 34: 9032–9037. https://doi.org/10.1073/pnas.1703220114.

Byers-Heinlein, K., E. Schott, A. M. Gonzalez-Barrero, et al. 2020. "MAPLE: A Multilingual Approach to Parent Language Estimates." *Bilingualism: Language and Cognition* 23, no. 5: 951–957. https://doi.org/10.1017/S1366728919000282.

Chai, J. H., M. McDonald, and E. Ko. 2024. *Investigating the Convergence of Child Language Assessment Measures in 14-Month-Old Korean Infants*. Glasgow, Scotland: Poster presented at the International Congress of Infant Studies.

Chater, N., S. M. McCauley, and M. H. Christiansen. 2016. "Language as Skill: Intertwining Comprehension and Production." *Journal of Memory and Language* 89: 244–254. https://doi.org/10.1016/j.jml.2015.11.004.

Creel, S. C. 2024. "Connecting the Tots: Strong Looking-Pointing Correlations in Preschoolers' Word Learning and Implications for Continuity in Language Development." *Child Development*. https://doi.org/10.1111/cdev.14157.

Cristia, A., M. Lavechin, C. Scaff, et al. 2021. "A Thorough Evaluation of the Language Environment Analysis (LENA) System." *Behavior Research Methods* 53, no. 2: 467–486. https://doi.org/10.3758/s13428-020-01393-5.

Dale, P. S., and M. Penfold. 2011. *Adaptations of the MacArthur-Bates CDI into Non*. U.S. English Languages. https://mb-cdi.stanford.edu/documents/AdaptationsSurvey7-5-11Web.pdf.

David, A., and L. Wei. 2008. "Individual Differences in the Lexical Development of French–English Bilingual Children." *International Journal of Bilingual Education and Bilingualism* 11, no. 5: 598–618. https://doi.org/10.1080/13670050802149200.

DeBolt, M. C., and L. M. Oakes. 2022. "Commentary on Six Solutions: Moving Forward With Measurement in Mind." *Infant and Child Development* 31, no. 5: e2324. https://doi.org/10.1002/icd.2324.

De Houwer, A., M. H. Bornstein, and D. B. Leach. 2005. "Assessing Early Communicative Ability: A Cross-Reporter Cumulative Score for the MacArthur CDI." *Journal of Child Language* 32, no. 4: 735–758. https://doi.org/10.1017/S0305000905007026.

Feldman, H. M., C. A. Dollaghan, T. F. Campbell, M. Kurs-Lasky, J. E. Janosky, and J. L. Paradise. 2000. "Measurement Properties of the MacArthur Communicative Development Inventories at Ages One and Two Years." *Child Development* 71, no. 2: 310–322. https://doi.org/10.1111/1467-8624.00146.

Fenson, L., P. S. Dale, J. S. Reznick, et al. 1994. "Variability in Early Communicative Development." *Monographs of the Society for Research in Child Development* 59, no. 5: i. https://doi.org/10.2307/1166093.

Fenson, L., P. S. Dale, J. S. Reznick, et al. 1993. *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual*. San Diego: Singular Publishing Group.

Fernald, A., V. A. Marchman, and A. Weisleder. 2013. "SES Differences in Language Processing Skill and Vocabulary Are Evident at 18 Months." *Developmental Science* 16, no. 2: 234–248. https://doi.org/10.1111/desc.12019.

Fernald, A., A. Perfors, and V. A. Marchman. 2006. "Picking up Speed in Understanding: Speech Processing Efficiency and Vocabulary Growth Across the 2nd Year." *Developmental Psychology* 42, no. 1: 98–116. https://doi.org/10.1037/0012-1649.42.1.98.

Fernald, A., R. Zangl, A. L. Portillo, and V. A. Marchman. 2008. "Looking While Listening: Using Eye Movements to Monitor Spoken Language Comprehension by Infants and Young Children." In *Language Acquisition and Language Disorders*, edited by I. A. Sekerina, E. M. Fernández, and H. Clahsen, Vol. 44, 97–135. John Benjamins Publishing Company. https://doi.org/10.1075/lald.44.06fer.

Fibla Reixachs, L. 2021. *Relating Language Input to Language Processes Early in Development: Using the Early Language Processing Task in UK and India* [Doctoral Dissertation]. University of East Anglia. School of Psychology. https://ueaeprints.uea.ac.uk/id/eprint/83017/.

Frank, M. C., M. Braginsky, D. Yurovsky, and V. A. Marchman. 2017. "Wordbank: An Open Repository for Developmental Vocabulary Data." *Journal of Child Language* 44, no. 3: 677–694. https://doi.org/10.1017/S0305000916000209.

Frank, M. C., M. Braginsky, D. Yurovsky, and V. A. Marchman. 2021. *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press. https://doi.org/10.7551/mitpress/11577.001.0001.

Ganek, H., and A. Eriks-Brophy. 2018. "Language ENvironment Analysis (LENA) System Investigation of Day Long Recordings in Children: A Literature Review." *Journal of Communication Disorders* 72: 77–85. https://doi.org/10.1016/j.jcomdis.2017.12.005.

Goodman, J. C., P. S. Dale, and P. Li. 2008. "Does Frequency Count? Parental Input and the Acquisition of Vocabulary." *Journal of Child Language* 35, no. 3: 515–531. https://doi.org/10.1017/S0305000907008641.

Hendrickson, K., S. Mitsven, D. Poulin-Dubois, P. Zesiger, and M. Friend. 2015. "Looking and Touching: What Extant Approaches Reveal About the Structure of Early Word Knowledge." *Developmental Science* 18, no. 5: 723–735. https://doi.org/10.1111/desc.12250.

Hessels, R. S., C. Kemner, C. Van Den Boomen, and I. T. C. Hooge. 2016. "The Area-Of-Interest Problem in Eyetracking Research: A Noise-Robust Solution for Face and Sparse Stimuli." *Behavior Research Methods* 48, no. 4: 1694–1712. https://doi.org/10.3758/s13428-015-0676-y.

Hirsh-Pasek, K., and R. M. Golinkoff. 1996. "The Intermodal Preferential Looking Paradigm: A Window onto Emerging Language Comprehension." In *Methods for Assessing Children's Syntax*, edited by D. McDaniel, C. McKee, and H. S. Cairns, 105–124. MIT Press.

Houston-Price, C., E. Mather, and E. Sakkalou. 2007. "Discrepancy Between Parental Reports of Infants' Receptive Vocabulary and Infants' Behaviour in a Preferential Looking Task." *Journal of Child Language* 34, no. 4: 701–724. https://doi.org/10.1017/S0305000907008124.

Hurtado, N., T. Grüter, V. A. Marchman, and A. Fernald. 2014. "Relative Language Exposure, Processing Efficiency and Vocabulary in Spanish–English Bilingual Toddlers." *Bilingualism: Language and Cognition* 17, no. 1: 189–202. https://doi.org/10.1017/S136672891300014X.

Hurtado, N., V. A. Marchman, and A. Fernald. 2007. "Spoken Word Recognition by Latino Children Learning Spanish as Their First Language." *Journal of Child Language* 34, no. 2: 227–249. https://doi.org/10.1017/S0305000906007896.

Kartushina, N., and J. Mayor. 2019. "Word Knowledge in Six- to Nine-Month-Old Norwegian Infants? Not Without Additional Frequency Cues." *Royal Society Open Science* 6, no. 9: 180711. https://doi.org/10.1098/rsos.180711.

Keenan, J. M., and B. MacWhinney. 1987. "Understanding the Relationship Between Comprehension and Production." In *Psycholinguistic Models of Production*, edited by H. W. Dechert and M. Raupach, Ablex Publishing Corporation. https://doi.org/10.1184/R1/6618917.v1.

Kremin, L. V., and K. Byers-Heinlein. 2021. "Why Not Both? Rethinking Categorical and Continuous Approaches to Bilingualism." *International Journal of Bilingualism* 25, no. 6: 1560–1575. https://doi.org/10.1177/13670069211031986.

Laing, C., and E. Bergelson. 2020. "From Babble to Words: Infants' Early Productions Match Words and Objects in Their Environment." *Cognitive Psychology* 122: 101308. https://doi.org/10.1016/j.cogpsych.2020.101308.

Law, F., and J. R. Edwards. 2015. "Effects of Vocabulary Size on Online Lexical Processing by Preschoolers." *Language Learning and Development* 11, no. 4: 331–355. https://doi.org/10.1080/15475441.2014.961066.

Law, J., and P. Roy. 2008. "Parental Report of Infant Language Skills: A Review of the Development and Application of the Communicative Development Inventories." *Child and Adolescent Mental Health* 13, no. 4: 198–206. https://doi.org/10.1111/j.1475-3588.2008.00503.x.

Legacy, J., P. Zesiger, M. Friend, and D. Poulin-Dubois. 2018. "Vocabulary Size and Speed of Word Recognition in Very Young French–English Bilinguals: A Longitudinal Study." *Bilingualism: Language and Cognition* 21, no. 1: 137–149. https://doi.org/10.1017/S1366728916000833.

Loi, E. C., V. A. Marchman, A. Fernald, and H. M. Feldman. 2017. "Using Eye Movements to Assess Language Comprehension in Toddlers Born Preterm and Full Term." *Journal of Pediatrics* 180: 124–129. https://doi.org/10.1016/j.jpeds.2016.10.004.

Marchman, V. A., V. N. Bermúdez, J. Y. Bang, and A. Fernald. 2020. "Off to a Good Start: Early Spanish-Language Processing Efficiency Supports Spanish- and English-language Outcomes at 4½ Years in Sequential Bilinguals." *Developmental Science* 23, no. 6: e12973. https://doi.org/10.1111/desc.12973.

Marchman, V. A., and A. Fernald. 2008. "Speed of Word Recognition and Vocabulary Knowledge in Infancy Predict Cognitive and Language Outcomes in Later Childhood." *Developmental Science* 11, no. 3: F9–F16. https://doi.org/10.1111/j.1467-7687.2008.00671.x.

Marchman, V. A., A. Fernald, and N. Hurtado. 2010. "How Vocabulary Size in Two Languages Relates to Efficiency in Spoken Word Recognition by Young Spanish–English Bilinguals." *Journal of Child Language* 37, no. 4: 817–840. https://doi.org/10.1017/S0305000909990055.

Marchman, V. A., E. C. Loi, K. A. Adams, M. Ashland, A. Fernald, and H. M. Feldman. 2018. "Speed of Language Comprehension at 18 Months Old Predicts School-Relevant Outcomes at 54 Months Old in Children Born Preterm." *Journal of Developmental and Behavioral Pediatrics* 39, no. 3: 246–253. https://doi.org/10.1097/DBP.0000000000000541.

Marchman, V. A., L. Z. Martínez, N. Hurtado, T. Grüter, and A. Fernald. 2017. "Caregiver Talk to Young Spanish-English Bilinguals: Comparing Direct Observation and Parent-Report Measures of Dual-Language Exposure." *Developmental Science* 20, no. 1: e12425. https://doi.org/10.1111/desc.12425.

Marchman, V. A., and C. Martínez-Sussmann. 2002. "Concurrent Validity of Caregiver/Parent Report Measures of Language for Children Who Are Learning Both English and Spanish." *Journal of Speech, Language, and Hearing Research* 45, no. 5: 983–997. https://doi.org/10.1044/1092-4388(2002/080.

Meylan, S. C., and E. Bergelson. 2022. "Learning through Processing: Toward an Integrated Approach to Early Word Learning." *Annual Review of Linguistics* 8, no. 1: 77–99. https://doi.org/10.1146/annurev-linguistics-031220-011146.

Mollica, F., and S. T. Piantadosi. 2017. "How Data Drive Early Word Learning: A Cross-Linguistic Waiting Time Analysis." *Open Mind* 1, no. 2: 67–77. https://doi.org/10.1162/OPMI_a_00006.

Moore, C., and E. Bergelson. 2022. “Examining the Roles of Regularity and Lexical Class in 18–26-Month-Olds’ Representations of How Words Sound.” *Journal of Memory and Language* 126: 1–17. https://doi.org/10.1016/j.jml.2022.104337.

Pearson, B. Z., S. C. Fernandez, V. Lewedeg, and D. K. Oller. 1997. “The Relation of Input Factors to Lexical Learning by Bilingual Infants.” *Applied PsychoLinguistics* 18, no. 1: 41–58. https://doi.org/10.1017/S0142716400009863.

Pearson, B. Z., S. C. Fernández, and D. K. Oller. 1993. “Lexical Development in Bilingual Infants and Toddlers: Comparison to Monolingual Norms.” *Language Learning* 43, no. 1: 93–120. https://doi.org/10.1111/j.1467-1770.1993.tb00174.x.

Peter, M. S., S. Durrant, A. Jessop, A. Bidgood, J. M. Pine, and C. F. Rowland. 2019. “Does Speed of Processing or Vocabulary Size Predict Later Language Growth in Toddlers?” *Cognitive Psychology* 115: 101238. https://doi.org/10.1016/j.cogpsych.2019.101238.

Pickering, M. J., and S. Garrod. 2013. “An Integrated Theory of Language Production and Comprehension.” *Behavioral and Brain Sciences* 36, no. 4: 329–347. https://doi.org/10.1017/S0140525X12001495.

Place, S., and E. Hoff. 2011. “Properties of Dual Language Exposure That Influence 2-year-olds’ Bilingual Proficiency.” *Child Development* 82, no. 6: 1834–1849. https://doi.org/10.1111/j.1467-8624.2011.01660.x.

Potter, C. E., and C. Lew-Williams. 2023. “Frequent vs. Infrequent Words Shape Toddlers' Real-Time Sentence Comprehension.” *Journal of Child Language*: 1–11. https://doi.org/10.1017/S0305000923000387.

Poulin-Dubois, D., E. Bialystok, A. Blaye, A. Polonia, and J. Yott. 2013. “Lexical Access and Vocabulary Development in Very Young Bilinguals.” *International Journal of Bilingualism* 17, no. 1: 57–70. https://doi.org/10.1177/1367006911431198.

Raykov, T., and G. A. Marcoulides. 2011. *Introduction to Psychometric Theory*. Routledge. https://doi.org/10.4324/9780203841624.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rescorla, L. A. 1980. “Overextension in Early Language Development.” *Journal of Child Language* 7, no. 2: 321–335. https://doi.org/10.1017/s0305000900002658.

Rocha-Hidalgo, J., and R. Barr. 2023. “Defining Bilingualism in Infancy and Toddlerhood: A Scoping Review.” *International Journal of Bilingualism* 27, no. 3: 253–274. https://doi.org/10.1177/13670069211069067.

Samuelson, L. K. 2021. “Toward a Precision Science of Word Learning: Understanding Individual Vocabulary Pathways.” *Child Development Perspectives* 15, no. 2: 117–124. https://doi.org/10.1111/cdep.12408.

Sander-Montant, A., M. López Pérez, and K. Byers-Heinlein. 2023. “The More They Hear the More They Learn? Using Data From Bilinguals to Test Models of Early Lexical Development.” *Cognition* 238: 105525. https://doi.org/10.1016/j.cognition.2023.105525.

Schott, E., C. Moore, and K. Byers-Heinlein. 2022. Banana and Banane: Cross-Language Phonological Overlap Supports Bilingual Toddlers’ Word Representations.”[Preprint], PsyArXiv. https://doi.org/10.31219/osf.io/hgdvq.

Schreiner, M. S., M. Zettersten, C. Bergmann, et al. 2024. Limited Evidence of Test-Retest Reliability in Infant-Directed Speech Preference in a Large Pre-Registered Infant Experiment.”[Preprint], PsyArXiv. https://doi.org/10.31234/osf.io/uwche.

Schwarz, G. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 6, no. 2. https://doi.org/10.1214/aos/1176344136.

Smolík, F., T. Sloupová, T. Fialová, K. Chládková, and N. Paillereau. 2023. *Parent-Reported Vocabulary and Looking-While-Listening in 164 Czech Toddlers: Support for Validity of Czech CDI: WG and CDI:WS Adaptations*. Boston, MA: Poster presented at the Boston University Conference on Language Development.

Spearman, C. 1904. “The Proof and Measurement of Association Between Two Things.” *American Journal of Psychology* 15, no. 1: 72. https://doi.org/10.2307/1412159.

Styles, S., and K. Plunkett. 2009. “What Is ‘Word Understanding’ for the Parent of a One-Year-Old? Matching the Difficulty of a Lexical Comprehension Task to Parental CDI Report.” *Journal of Child Language* 36, no. 4: 895–908. https://doi.org/10.1017/S0305000908009264.

Swingley, D. 2011. “The Looking-While-Listening Procedure.” In *Research Methods in Child Language*, edited by E. Hoff, 1st ed., 29–42. Wiley. https://doi.org/10.1002/9781444344035.ch3.

Trudeau, N., H. Frank, and D. Poulin-Dubois. 1999. “Une adaptation en français Québécois du MacArthur Communicative Development Inventory.” *La Revue d’orthophonie et d’audiologie* 23, no. 2: 61–73.

Vihman, M. M., and L. McCune. 1994. “When Is a Word a Word?” *Journal of Child Language* 21, no. 3: 517–542. https://doi.org/10.1017/S0305000900009442.

Weaver, H., and J. Saffran. 2024. *Word-Level Convergent Validity Between Looking-While-Listening and Caregiver Report of Vocabulary*. Glasgow, Scotland: Poster Presented at the International Congress of Infant Studies.

Weisleder, A., and A. Fernald. 2013. “Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary.” *Psychological Science* 24, no. 11: 2143–2152. https://doi.org/10.1177/0956797613488145.

Weisleder, A., M. Friend, A. Sin Mei Tsui, and V. A. Marchman. 2024. “Using Parent Report to Measure Vocabulary in Young Bilingual Children: A Scoping Review.” *Language Learning* 74, no. 2: 468–505. https://doi.org/10.1111/lang.12617.

Werker, J. F., C. T. Fennell, K. M. Corcoran, and C. L. Stager. 2002. “Infants’ Ability to Learn Phonetically Similar Words: Effects of Age and Vocabulary Size.” *Infancy* 3, no. 1: 1–30. https://doi.org/10.1207/S15327078IN0301_1.

White, K. S., and J. L. Morgan. 2008. “Sub-Segmental Detail in Early Lexical Representations.” *Journal of Memory and Language* 59, no. 1: 114–132. https://doi.org/10.1016/j.jml.2008.03.001.

Yoder, P. J., S. F. Warren, and H. A. Biggar. 1997. “Stability of Maternal Reports of Lexical Comprehension in Very Young Children With Developmental Delays.” *American Journal of Speech-Language Pathology* 6, no. 1: 59–64. https://doi.org/10.1044/1058-0360.0601.59.

Zangl, R., L. Klarman, D. Thal, A. Fernald, and E. Bates. 2005. “Dynamics of Word Comprehension in Infancy: Developments in Timing, Accuracy, and Resistance to Acoustic Degradation.” *Journal of Cognition and Development* 6, no. 2: 179–208. https://doi.org/10.1207/s15327647jcd0602_2.

Zettersten, M., C. Bergey, N. S. Bhatt, et al. 2021. Peekbank: Exploring Children’s Word Recognition through an Open, Large-Scale Repository for Developmental Eye-Tracking Data.”[Preprint], PsyArXiv. https://doi.org/10.31234/osf.io/ep693.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.