*Developmental Science*

WILEY

Developmental Science

RESEARCH ARTICLE OPEN ACCESS

# Interrogating Early Word Knowledge: Factors That Influence the Alignment Between Caregiver-Report and Experimental Measures

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

[1]Department of Psychology, University of Wisconsin–Madison, Madison, USA | [2]Department of Speech, Language & Hearing Sciences, Boston University, Boston, USA

**Correspondence:** Haley Weaver (hjweaver@bu.edu)

**Received:** 19 April 2025 | **Revised:** 29 September 2025 | **Accepted:** 19 October 2025

**Keywords:** caregiver report | eye-gaze | infants | looking-while-listening | MB-CDI | word recognition

## ABSTRACT

Questions about early word knowledge pervade the literature on both typical and atypical language trajectories. To determine which words an infant knows, researchers have relied on two types of measures—caregiver-report and eye-gaze behavior. When these measures are compared, however, their results frequently fail to converge, making it difficult to ascertain whether a given infant knows a given set of words. What are the reasons for these misalignments in gold-standard tasks that are designed to investigate the same underlying construct, and can convergent validity be improved? The current study was designed to investigate multiple methodological features of caregiver-report and looking-while-listening (LWL) tasks hypothesized to contribute to their alignment. American English-learning infants (18–20 months; $N = 52$) completed an LWL task assessing their understanding of eight early-acquired words. Caregivers reported their child's knowledge of the same eight words, as well as their confidence in their responses and the amount of time they spend with their child. Overall, caregivers' reports of word knowledge did not predict infants' eye-gaze behavior. However, the measures were more likely to be aligned when caregivers reported higher confidence in their responses. Caregivers' reports about both the target and the distractor word on each trial were related to infants' eye-gaze behavior, suggesting that LWL tasks capture knowledge about the labels of both objects tested, not just the label of the target object. The results suggest several critical methodological modifications that could be implemented to improve the measurement validity of both caregiver-report and eye-gaze measures of word comprehension.

## 1 | Introduction

Assessing infants' knowledge about specific words poses unique challenges. Infants have limited language production skills and often understand many more words than they can say. Thus, the typical methods used to measure lexical knowledge in adults, such as asking participants to report features of known words (e.g., McRae et al. 1997), are not feasible in infant populations. Instead, indirect measures of word knowledge, such as caregiver reports of vocabulary (e.g., MacArthur-Bates Communicative Development Inventories [MB-CDI]; Fenson et al. 2007) and infants' eye-gaze behavior (e.g., looking-while-listening [LWL]; Fernald et al. 2008) are widely used to assess early vocabulary.

Caregiver-report measures have many methodological advantages and, as such, have been widely used to investigate early vocabulary. For example, vocabulary checklists such as the MB-CDI are simple and cost-effective to administer. They provide a wealth of data that allows researchers to glean vocabulary information beyond what is observable in a brief infant study. Furthermore, these checklists have high concurrent and predictive validity with other measures of language development

**Summary**

- This study provides novel insights to improve the validity of infant vocabulary measurement by highlighting factors that explain differences between caregivers' reports and eye-gaze behavior.
- Caregivers' reports of individual word knowledge failed to converge with infants' eye-gaze behavior during a looking-while-listening (LWL) task.
- Caregivers' reports of individual word knowledge aligned more strongly with their infant's eye-gaze behavior when caregivers reported more confidence in their responses.
- Caregivers' reports of both the target and the distractor words predicted infants' eye gaze behavior.

(Feldman et al. 2005; Fernald et al. 2006; Fernald and Marchman 2012; Marchman and Fernald 2008; Nordahl-Hansen et al. 2014; Ring and Fenson 2000; Sachse and Von Suchodoletz 2008). Access to MB-CDI datasets like WordBank has allowed researchers to quantify the size, variability, and semantic content of early lexicons (e.g., Frank et al. 2021; Marchman et al. 2023; Peters and Borovsky 2019).

Caregiver-report methods also pose several challenges to measuring knowledge about specific words. Caregivers must make categorical decisions about word knowledge (e.g., does not understand, only understands, or understands and says) based on cumulative language experiences and shared interactions with their child. For words that are not yet said, caregivers must interpret their infant's non-linguistic behaviors as an indication of understanding. Notably, these discrete responses do not allow caregivers to provide insight into their child's partial word knowledge (McMurray 2007; Swingley 2010). For example, a child may correctly use a word like "apple" to refer to apple slices, but may also overextend this word to refer to other round fruits such as oranges. These kinds of experiences may make caregivers less confident that their child knows a word. Because uncertainty and graded knowledge are not captured by caregiver-report measures, all-or-nothing checklists can feel quite daunting for caregivers to complete, given the extent and nature of their interactions with their children.

Another possible limitation in vocabulary checklist methods is that caregivers vary in the number of experiences they can draw on as evidence that their child knows a word. For example, caregivers who primarily stay at home with their infants may have more opportunities to observe their child's language than other caregivers. These differences likely contribute to increased variability when comparing vocabulary sizes across infants (Frank et al. 2021), as well as inconsistency across reporting sessions for individual words (Arunachalam et al. 2022). While some caregivers may be able to veridically report their child's level of understanding of some words, others may be more uncertain about which words their child knows.

A potentially more direct measure of word comprehension is eye-gaze behavior. In preferential looking tasks, such as LWL (Fernald et al. 2008) and intermodal preferential looking (Golinkoff et al. 2013; Hirsh-Pasek et al. 1987), infants view two images on a screen and hear speech labeling one of the images. Notably, these tasks allow researchers to measure continuous differences in infants' word knowledge, as indexed by the speed or accuracy of word recognition. Gaze-behavior methods have revealed an impressive range of early comprehension abilities, often earlier than suggested by caregivers' reports (e.g., Bergelson and Swingley 2012, 2015; Campbell and Hall 2022; Garrison et al. 2020).

Importantly, methodological decisions influence researchers' ability to measure whether infants know a word using gaze behavior (Kucker and Lorenz 2025). For example, word recognition performance is affected by the visual salience of the stimuli (Pomper and Saffran 2019), the frequency of the target word (Kartushina and Mayor 2019; Potter and Lew-Williams 2023), and the specific images selected as targets (Casey et al. 2023). These studies suggest that stimulus selection is a critical design feature. Indeed, conclusions about a child's word comprehension based solely on gaze behavior are limited to the precise target images tested in the context of the specific distractors with which they were paired. Thus, despite the benefits of using a more direct measure of word comprehension, the experimental designs typically used in LWL studies may make it difficult to definitively determine whether a given infant understands a given word.

Word recognition performance may also reflect several different underlying processes. For instance, target looking may reflect knowledge of an explicit mapping between a label and its referent. This word knowledge may also be graded; infants with more robust mappings may be faster and more accurate at identifying a target than those with weaker mappings (Hendrickson et al. 2015, Hendrickson et al. 2017; Smolak et al. 2021). Alternatively, target looking may be the result of knowing the label for the distractor object, rather than comprehending the target label, reflecting a mutual exclusivity bias (Markman 1990; Markman and Wachtel 1988). Using this strategy, infants as young as 17 months of age could leverage their knowledge of the distractor object (and its label) to eliminate it as a possible referent for an unfamiliar word (e.g., Byers-Heinlein and Werker 2009; Halberda 2003; Houston-Price et al. 2010). For instance, upon hearing "bunny," an infant may look more at an image of a rabbit not because they know its label, but because it is paired with a familiar distractor object (e.g., sock) for which they already know the label. While this mutual exclusivity strategy would result in more target looking, this behavior likely reflects knowledge about the label of the distractor object rather than the target word. All of these considerations suggest that word recognition accuracy may be the result of many different kinds of word knowledge.

Given the limitations of both indirect methods (caregiver report of vocabulary) and direct measures (eye-gaze behavior), what is the most valid method for determining which words a given infant knows? Several prior studies have attempted to establish the convergent validity of LWL and caregiver-report methods both averaged across words and for individual words (Fernald et al. 2006; Houston-Price et al. 2007; López Pérez et al. 2025; Lorang et al. 2023; Marchman and Fernald 2008; Styles and Plunkett 2009; Syrnyk and Meints 2017; Venker et al. 2016). To date, there is mixed evidence that the two methods converge. With respect to overall vocabularies, overall vocabulary size is less strongly associated with infants' word recognition abilities

at certain ages (e.g., young infants; Bergelson and Swingley 2012, 2015; Fernald et al. 2006). At the same time, the two methods are strongly associated across developmental time (Fernald et al. 2006; Fernald and Marchman 2012; Marchman and Fernald 2008; Weisleder and Fernald 2013). When considering individual word knowledge, some studies demonstrate convergent validity (Styles and Plunkett 2009; Syrnyk and Meints 2017); others suggest that caregivers' reports underestimate word comprehension both in typically-developing (Bergelson and Swingley 2012; Houston-Price et al. 2007) and neurodiverse populations (Arunachalam et al. 2022; Lorang et al. 2023; Venker et al. 2016).

One possible explanation for the lack of consensus regarding convergent validity is that prior studies have not directly examined the design issues that may influence alignment between the two methods. Some caregivers may underestimate word knowledge if they are uncertain about whether their infant's prior behaviors count as evidence for understanding or saying a word. On the other hand, some caregivers may be overzealous about their child's language (e.g., interpreting babbling sounds as producing a word). At the same time, the LWL task may not veridically measure infants' word understanding. The forced-choice structure of an LWL task means that in some cases, knowledge of the distractor object may lead infants to appear to understand a target word when they actually do not. However, this may be less likely to occur in naturalistic contexts (e.g., an infant may be asked to bring their caregiver a tiger, which is amongst several toys of varying familiarity). The possibility that mutual exclusivity effects may be more likely to occur in LWL tasks than in children's home environments may also impact alignment between eye-gaze tasks and caregiver-report measures. Finally, differences may arise from the fact that caregivers' reports are all-or-none, whereas eye-gaze behavior is continuous. Partial knowledge about the meaning of a word, as evidenced by above-chance target looking (e.g., 65% accuracy over an analysis window), may not be evident to caregivers interacting with the child, leading to misalignments between the two measurements.

Prior studies examining convergent validity have either focused on caregivers' uncertainty (Hanania 2021; Lorang et al. 2023) or LWL stimuli (Styles and Plunkett 2009) as an explanation for low convergent validity. These studies have separately demonstrated that caregivers are more certain when reporting words that are said (Hanania 2021; Lorang et al. 2023) and that the two methods align when the LWL image pairs are semantically related (e.g., cup and bowl; Styles and Plunkett 2009). Although these findings provide compelling hints about the factors that influence alignment, they necessarily assume that one of the methods is a more veridical measure of word knowledge. Some studies argue that LWL is more accurate and therefore identify potential improvements in caregiver-report measures (Hanania 2021; Houston-Price et al. 2007; Lorang et al. 2023). Others consider caregiver reports to be a stronger measure of word knowledge, consequently modifying the LWL task to better approximate those reports (Styles and Plunkett 2009; Syrnyk and Meints 2017).

Crucially, neither method provides "ground truth." Each method has limitations, and it is unclear how these limitations jointly impact convergent validity. Therefore, in the current study, we examined the potential moderating factors of each method in tandem to investigate whether and how alignment between caregivers' reports and LWL behavior for specific words can be improved. To do so, we (1) attempted to replicate associations between productive vocabulary (per caregivers' reports) and LWL performance averaged across words, (2) investigated convergent validity for individual word knowledge, and (3) examined moderators of convergent validity (i.e., confidence, the time a caregiver spends with their infants, and LWL image pairs). To test these questions, we adapted both the LWL task and the caregiver-report instruments in ways that we thought could increase our ability to align the two measures.

The LWL task included multiple referent images for each word and paired each target with multiple different distractor items. While most studies in the infant word recognition literature use yoked stimuli (e.g., *cup* is always paired with *shoe*) due to benefits in terms of experimental control, they also leave the results more vulnerable to mutual exclusivity effects. For example, if looking behavior for both members of the pair (*cup* and *shoe)* is above chance, it's not possible to determine whether the child actually knows both words or if, instead, they know just one member of the pair and could use mutual exclusivity to guess the other. Unyoked stimuli, however, have more ecologically valid variability, better match the task that caregivers are asked to do when reporting vocabulary (e.g., consider their child's wealth of experiences with different objects), and, therefore, potentially improve convergent validity. This design also afforded us an opportunity to directly investigate whether the distractor images influenced alignment between the two measures, as implicated in prior studies (Pomper and Saffran 2019; Styles and Plunkett 2009). In particular, we could conduct exploratory analyses investigating potential mutual exclusivity and age-of-acquisition effects driven by infants' knowledge of the labels for the distractor objects.

Finally, the caregiver-report tasks included the MB-CDI and a study-specific vocabulary checklist for the words tested in the LWL task. These checklists also included a measure of caregivers' confidence in their reports of word knowledge. Confidence was measured using a single question at the end of the MB-CDI, and was measured for each of the words on the study-specific checklist. Based on previous findings, we expected that caregivers' reports of word knowledge would better reflect their child's LWL behavior when a caregiver was more confident in their report (Lorang et al. 2023). In addition, we examined whether the amount of time the reporting caregiver spends with their child moderated convergent validity. We predicted that caregivers who spend more time with their child may also have more opportunities to observe their child's word knowledge, thus improving alignment between the two measures. The methods and analysis plan were preregistered on OSF: https://osf.io/c8zws.

## 2 | Method

### 2.1 | Participants

The final sample included 52 infants between 18 and 20 months ($M_{age}$ = 18.81, 28 females) from a mid-size US city. We chose this age because prior studies demonstrated mixed evidence for alignment between eye-gaze behavior and caregiver report of vocabulary in this age span (Houston-Price et al. 2007; López

**TABLE 1** | MB-CDI production and comprehension norms for target nouns.

| Noun | Proportion understanding | Proportion producing | | |
|---|---|---|---|---|
| | 18-months | 18-months | 19-months | 20-months |
| Ball | 0.97 | 0.84 | 0.89 | 0.93 |
| Apple | 0.82 | 0.49 | 0.61 | 0.61 |
| Sock | 0.86 | 0.38 | 0.48 | 0.48 |
| Flower | 0.67 | 0.28 | 0.44 | 0.47 |
| Toothbrush | 0.81 | 0.23 | 0.36 | 0.37 |
| Peas | 0.47 | 0.14 | 0.26 | 0.26 |
| Glasses | 0.52 | 0.14 | 0.25 | 0.26 |
| Crayon | Unknown | 0.12 | 0.26 | 0.25 |

*Note:* The table shows the proportion of infants (18- to 20-months) reported to understand and produce the eight target words according to WordBank (Frank et al. 2017).

Pérez et al. 2025; Styles and Plunkett 2009). The preregistered sample size was based on a priori power analysis in G*Power (Version 3.1.9.6; Erdfelder et al. 1996) using a medium effect size. Previous studies report a large effect for the interaction between reported word knowledge and analysis window (i.e., behavior prior to labeling vs. behavior following the label; Houston-Price et al. 2007; Styles and Plunkett 2009). We powered our sample to a slightly smaller effect to account for differences in analytic strategies and publication bias. The preregistered sample size provides 90% power to detect an effect as small as $d = 0.3$. Fourteen additional infants participated but were excluded from analyses due to insufficient data contribution ($N = 10$; contributing fewer than eight test trials) or equipment failure ($N = 4$).

Infants were reported to be born full term, have no history of developmental disabilities, have normal hearing and vision, and hear English at least 90% of the time. Of the caregivers who completed a demographics survey ($N = 43$), most reported coming from upper-middle-class homes; 95% of caregivers held a bachelor's degree or higher, and 93% had a household income of at least $90,000. The sample was predominantly White (83%); 15% of the sample identified as belonging to multiple racial categories, and 2% identified as Asian. No one identified as Hispanic.

## 2.2 | Materials

We selected eight nouns from the MB-CDI Words and Sentences form. To ensure variance in the dataset, the words were chosen to vary in age of acquisition (AoA) as indexed by the proportion of 18- to 20-month-olds reported to understand them (see Table 1; Frank et al. 2017). We used four different images for each noun that varied on several features, including color and shape (see Figure 1a). Images were edited and placed on a 550 × 550 pixel grey background.

A female speaker of the local dialect recorded sentences including the target words using infant-directed speech. Each target word appeared in four frames: "Where's the [noun]," "Find the [noun]," "Do you see the [noun]?," "Look at the [noun]." An attention-getting phrase ("Check that out!" or "Wow, that's cool!") followed the target word to encourage infants to continue to fixate the screen. All sentences were edited to have the same intensity (67 dB) and duration (7790 ms), with the onset of the noun occurring after 2850 ms.

## 2.3 | Procedure

After providing informed consent, caregivers completed a survey that included a study-specific vocabulary checklist and information about their child's home environment. Caregivers reported whether their child "did not understand," "only understood," or "understood and said" each of the eight words from the LWL task. For each word, caregivers provided a confidence rating using a scale from 0 (completely unsure) to 100 (very sure). Caregivers also reported the number of hours spent with their child on a typical day and several demographic factors, including income, education level, and race.

Participants were then seated in a soundproof booth equipped with a video camera, a 55" screen, and a Tobii x60 eye tracker. Eye-gaze was calibrated prior to the LWL procedure using a five-point procedure. Sessions were videorecorded to allow for manual gaze-coding if automatic eye tracking was unsuccessful.

Infants completed an LWL task that tested their recognition of eight nouns: apple, ball, crayon, flower, glasses, peas, sock, and toothbrush. On each of 32 trials, infants saw two images, one on either side of the screen, and heard a sentence labeling one of the images. Importantly, there were four different target images for each word, and the same target-distractor pair never occurred twice (see Figure 1b for an example). All images occurred equally often as targets and distractors. There were attention-getting images after every fourth trial.

In addition to the survey and the LWL task, caregivers completed a modified version of the MB-CDI short form level IIB. Typically, this form is a 100-word productive vocabulary checklist. However, we modified the form to add comprehension questions for each item, as well as a single question regarding their global confidence in completing the form (0–100).


**FIGURE 1** | The top panel (a) displays each of the four images used to test the target nouns. The bottom panel (b) depicts an example of the target-distractor pairings for *ball*.

## 2.4 | Measures

### 2.4.1 | Survey Data

To assess whether caregivers' reports and LWL behavior are better aligned when caregivers spend more time with their child, we computed the proportion of time the caregiver completing the form spent with their infant (computed separately and combined for weekdays and weekends) as a function of the typical number of waking hours for each child.

### 2.4.2 | Gaze Behavior

To assess infants' overall accuracy, we computed the baseline-corrected proportion of target looking for each participant, averaged across all 32 trials. For each infant, we calculated the proportion of target fixations during a critical window of 300 to 2000 ms following noun onset (Garrison et al. 2020) and subtracted the proportion of target fixations during a baseline window prior to noun onset (i.e., −2000 to 0 ms; Bergelson and Swingley 2012; Garrison et al. 2020; Weaver et al. 2024). Baseline correction has been used in prior LWL studies to account for salience differences between the target and distractor images (Bergelson and Swingley 2012; Garrison et al. 2020; Weaver et al. 2024). The literature looking at alignment between caregiver-report and LWL measures often included an additional parameter for pre- and post-naming (Styles and Plunkett 2009); however, baseline correction provides a conceptually equivalent approach to this analytic procedure without introducing additional model complexity.

To measure word-level accuracy, we calculated infants' baseline-corrected proportion of target looking for each trial in the LWL task. Word-level accuracy reflects the proportion of time infants spent fixating the target image out of the total time they spent fixating either image relative to their target fixations prior to labelling.

The third outcome variable was reaction time (RT), which we defined as the latency to first fixate the target image (Fernald et al. 2008). RTs were only computed on trials in which infants were fixating the distracter image at noun onset ($N = 598$ trials).

Trials with too much missing data (less than 50% of fixations to either the left or the right side of the screen during the baseline or the critical window) were not included in the analyses. Four hundred (out of a possible 1606) trials met criteria for exclusion. Of the excluded trials, 6.79% were missing data in the baseline window, 7.16% of the excluded trials were missing data in the critical window, and 10.96% of the excluded trials were missing data in both windows.

### 2.4.3 | CaregiverReport of Vocabulary

The MB-CDI provided a measure of infants' overall vocabulary knowledge. Using the MB-CDI norms for both male and female infants, we calculated each infant's productive vocabulary percentile (Marchman et al. 2023). We used vocabulary production percentile as our primary measure because it has been used to examine the convergent validity of the MB-CDI and other language measures (Feldman et al. 2005).

The caregiver survey provided an additional measure of word knowledge that was specific to the words we tested in the LWL task. We defined infants' study-specific vocabulary size as the number of words (out of 8) a caregiver reported their child to say. Again, we used production because it has been used to examine convergent validity.

## 3 | Results

All data processing and analyses were conducted in R (R Core Team, 2024; version 4.4.2). Linear models were fit using the

stats package (version 4.4.2), and linear-mixed effects models were fit using the lme4 package (version 1.1.35.3, Bates et al. 2015). We computed $p$ values using the Anova() function, which uses Kenward–Roger approximation to estimate degrees of freedom for $F$-tests. For each model, variables were centered unless otherwise specified. Linear-mixed effects models were fit using the maximal random effects structure and pruned iteratively following recommendations by Brauer and Curtin (2018) and Muradoglu et al. (2023). In what follows, we report the final models that converged. A full list of the model specifications can be found in the Supporting Information (Table S1). The data and analytic scripts supporting the results can be found on OSF: https://osf.io/xgnvk/overview?view_only=4c179810233e4066827ade18ba01a15c.

### 3.1 | Overall Word Knowledge

Caregivers reported on the MB-CDI that infants produced an average of 19.89 words (SD = 14.63, range = 3–62), which corresponds to an average MB-CDI percentile score of 27.17 (range = 4–80). They were reported to understand 27.14 words on average (SD = 11.93, range = 7–55); percentiles are not available for comprehension in this age group. With respect to the tested target words, caregivers reported that infants produced an average of 2.70 (range = 1–7) and comprehended an average of 4.09 (range = 1–8) out of the 8 words. For the LWL task, infants' average accuracy was 58.18% (SD = 0.345; baseline-corrected$_M$ = 0.088, baseline-corrected$_{SD}$ = 0.421). As a group, they recognized 4 out of 8 words significantly above chance (i.e., "apple," "ball," "glasses," and "sock").

The following two analyses addressed the alignment between caregivers reports and LWL for word knowledge in general. First, we investigated whether MB-CDI scores predicted accuracy on the LWL task by fitting a linear model regressing our measure of overall accuracy (average baseline-corrected proportion of target looking) on overall vocabulary knowledge (MB-CDI production percentile score). As can be seen in Figure 2, productive vocabulary percentile was not a significant predictor of LWL accuracy ($b$ = 0.0001, $F(1, 41) = 0.028$, $p = 0.866$). The same pattern of results holds for two other MB-CDI measures: raw productive vocabulary ($p = 0.97$) and total comprehension vocabulary size ($p = 0.22$).

Next, we explored whether caregivers' reports about the specific words used in our task predicted overall LWL accuracy. We fit a linear model regressing our measure of overall accuracy on study-specific vocabulary size. This model tested whether infants who produced more of the 8 target words performed more accurately overall on the LWL task. Study-specific vocabulary size was not a significant predictor of LWL behavior ($b = 0.079$, $F(1, 43) = 0.915$, $p = 0.344$). Alternative measures of study-specific vocabulary size were also not associated with LWL behavior: number of words produced ($p = 0.599$), and number of words comprehended ($p = 0.970$).

We also examined relations between RT and overall vocabulary score. We fit analogous models to those above (i.e., MB-CDI productive percentile score and proportion of the study-specific words produced) using RT as the outcome variable. Neither measure predicted RT ($b = -0.682$, $F(1, 41) = 0.215$, $p = 0.645$; $b = -251.8$, $F(1, 43) = 3.561$, $p = 0.066$).

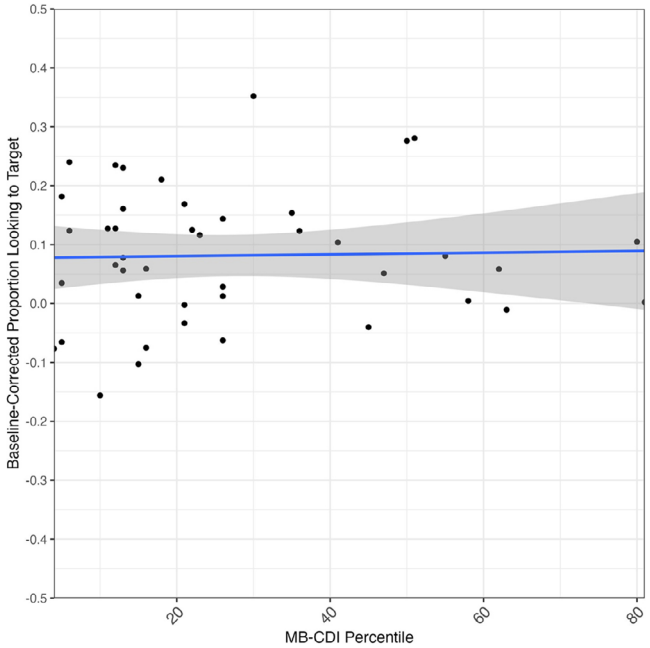


**FIGURE 2** | Baseline-corrected proportion of target looking as a function of productive vocabulary size. Data points reflect participants' baseline-corrected proportion of target looking averaged across trials and words. Error bars reflect standard error of the point estimate.

### 3.2 | Specific Word Knowledge

The primary analyses of interest investigated whether infants' recognition of specific words on the LWL task aligned with caregivers' report of their child's knowledge about those words. Caregivers' report of word knowledge was coded using contrast codes (i.e., −0.5 for words reported as unknown and 0.5 for words reported as understood or said) in the first set of analyses, and coded using linear contrasts (i.e., −1 for unknown, 0 for understands, and 1 for says) in the second set of analyses. We used both coding schemes because they represent theoretically different interpretations. In the model using binary contrast codes, we are assessing whether caregivers' report of any kind of word comprehension is correlated with word recognition. In the model using linear contrasts, we are testing whether there are distinct increases in word recognition performance depending on the level of word knowledge that was reported.

We first assessed alignment between LWL behavior and caregivers' reports of individual words using binary contrast codes. We fit an LMEM regressing word-level accuracy (baseline-corrected proportion of target looking on each trial) on caregiver report of word knowledge, including by-subject and by-item random intercepts. On average, infants recognized the target words significantly above chance ($b = 0.065$, $F(1, 15.96) = 8.618$, $p = 0.010$). Caregiver report of word knowledge, however, was not a significant predictor of word recognition accuracy (Figure 3; $b = 0.064$, $F(1, 352.37) = 3.502$, $p = 0.062$). That is, infants who were reported to understand or say a specific word did not look significantly more at the target image after hearing that word than

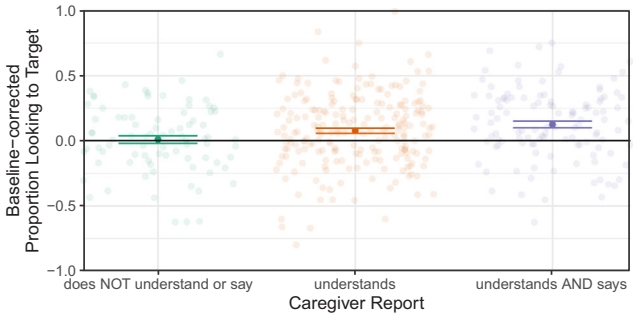


**FIGURE 3** | Baseline-corrected proportion of target looking as a function of caregiver report for individual words. Each data point represents a participant's baseline-corrected proportion of target looking for a given word. Error bars were calculated as standard error of the point estimate.

infants who were reported to not know the word. An analogous model using RT as the primary measure of LWL behavior revealed similar results ($b = -46.08$, $F(1, 3.651) = 0.478$, $p = 0.531$).

We next assessed alignment between LWL behavior and caregivers' reports of individual words using the linear contrast codes. We regressed word-level accuracy on linear (−0.5 for unknown, 0 for comprehends, and 0.5 for says) contrasts for caregiver report of word knowledge, including a by-subject random slope and a by-item random slope and intercept for caregiver report. Following recommendations from Abelson and Prentice (1997), we also included contrast codes for a quadratic effect of caregiver report (0.5 for unknown, −1 for comprehends, and 0.5 for says) and a by-subject random slope for this effect to ensure that the linear pattern fits the data better than an alternative relation. The linear effect for caregiver report of word knowledge was not a significant predictor of word-level accuracy ($b = 0.076$, $F(1, 6.153) = 5.387$, $p = 0.058$) nor was the quadratic effect ($b = -0.009$, $F(1, 49.763) = 0.105$, $p = 0.748$). Therefore, infants who were reported to say a word ($M = 0.137$, SD = 0.393) were not significantly more accurate at recognizing words than infants who were only reported to comprehend a word ($M = 0.083$, SD = 0.431), and infants who were reported to only comprehend a word were not significantly more accurate at recognizing words than infants who were reported to not know a word ($M = 0.027$, SD = 0.433). In a parallel model using RT as the primary measure of word recognition, we similarly found no significant relation between the linear effect of caregiver report and RT ($b = -50.15$, $F(1, 12.028) = 0.652$, $p = 0.435$).

### 3.3 | Confidence

We hypothesized that caregivers' confidence in their report of word knowledge would moderate the relation between caregiver report and word recognition. That is, caregivers who are more confident in their responses may also be more aligned with their child's eye-gaze behavior. We examined this question with respect to caregivers' reports of both the overall vocabulary score and caregivers' reports of specific words.

To examine the effect of confidence on the relation between overall caregiver-reported vocabulary and LWL accuracy, we regressed infants' overall accuracy on MB-CDI percentile, confidence, and their interaction. MB-CDI percentile, overall confidence, and the interaction between confidence and MB-CDI percentile were not significant predictors of overall accuracy ($p = 0.900$, $p = 0.653$, and $p = 0.738$, respectively). Importantly, however, caregivers reported relatively high confidence in their report of their child's vocabulary ($M = 82.23$; SD = 11.056).

Next, we examined whether confidence moderated alignment between the two measures for specific words. Caregivers' average confidence for their response for a given word was high ($M = 83.81\%$; SD = 20.17). Confidence varied across caregiver-reported word knowledge. On average, caregivers reported being most confident when reporting words their child was saying ($M = 96.39\%$; SD = 9.90), followed by words their child only understood ($M = 79.86\%$; SD = 19.16) and words their child did not know ($M = 74.47\%$; SD = 25.39). To examine relations between caregiver report of word knowledge, gaze behavior, and confidence, we fit an LMEM regressing infants' word-level accuracy on caregivers' reports of those same words (coded −0.5 for "does not understand" and 0.5 for "understands and/or says"), confidence about their response for each word, and their interaction, including a by-subject random slope for caregiver report and a by-item random slope and intercept for the interaction between caregiver report and confidence. Caregiver report of vocabulary, confidence about each word, and their interaction were not significant predictors of word-level accuracy ($p = 0.236$, $p = 0.664$, $p = 0.277$).

We preregistered a linear regression to assess moderator effects. However, upon further reflection, we realized that a linear regression presupposes that the relationship between confidence and LWL accuracy is linear. That is, the model tests the hypothesis that infants are more accurate on the LWL task when their caregivers are more confident in their response. However, this is not necessarily the case. A parent could be very confident that their child does not understand a word, for example, which could result in a negative relationship between confidence (high) and accuracy (chance). Instead, the more likely relationship concerns the relationship between confidence and alignment. We would expect a parent who is more confident to be better aligned with their child's behavior, both by knowing which words their child knows (associated with higher accuracy on the LWL task) and which words their child does not know (associated with lower accuracy on the LWL task).

To test this hypothesis, we ran an alternative exploratory analysis using alignment as the outcome variable rather than accuracy. Each word was given an alignment score that reflects whether a caregiver's report of word knowledge aligned with their child's behavior on the LWL task (Table 2). First, we converted accuracy to a binary measure, which reflected whether a child recognized each word on the LWL task. To do so, we calculated a 95% confidence interval (CI) around chance for each child using the Clopper–Person exact method from the binom package (Clopper and Pearson 1934; Dorai-Raj 2022). We reasoned that a 95% CI would allow us capture whether looking behavior was at chance (i.e., 50%) or at values equivalent to chance (e.g., 51%). LWL accuracies above the 95% CI were considered understanding a word, while values at or below the 95% CI were considered as not understood. Then, we compared infants' binary accuracy to their caregiver's report of word knowledge to determine alignment. As an example, a word would receive an alignment score of one if the

**TABLE 2** | Alignment scores.

| LWL | Caregiver report | Alignment |
|---|---|---|
| Below or within the 95% CI around chance | Does not understand/say | 1 |
| | Understands or says | 0 |
| Above the 95% CI around chance | Does not understand/say | 0 |
| | Understands or says | 1 |

*Note:* The table shows the alignment coding for LWL and caregiver-report data. Chance performance is defined as the 95% CI around 50% because LWL uses a two-alternative forced choice task. Thus, a score above chance indicates increased looking to the target relative to baseline.

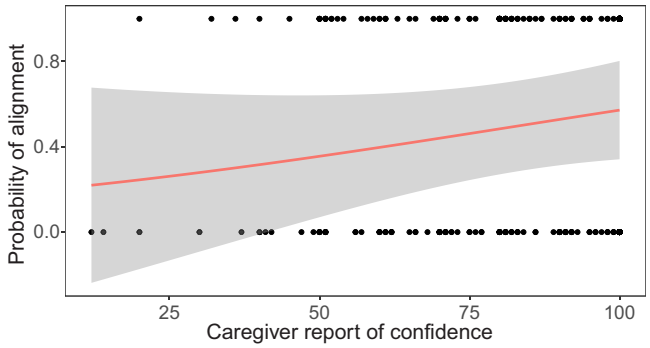


**FIGURE 4** | Probability of alignment depending on caregivers' confidence about their report of word knowledge. Dots represent an individual participant's alignment (0 or 1) for a given word. The red line was created using model fits, and the error bands reflect standard error of the point estimate.

caregiver reported the word as understood or said, and the child's LWL accuracy for that word was above the 95% CI. Similarly, we would assign an alignment score of one if a caregiver reported that their child did not understand a word and the child's LWL accuracy for that word was inside the 95% CI or lower than the lower bound of the CI. Figure S1 depicts word-level LWL accuracy and caregiver report of word knowledge for each subject.

Using alignment as the outcome measure, we fit a generalized linear mixed effects model predicting alignment from confidence (centered), including a by-subject random intercept and a by-item random slope for confidence. Confidence was a significant predictor of alignment ($b = 0.018$, 95% CI [0.004, 0.032], $p = 0.009$, OR = 1.018). Therefore, the word-level data from caregivers and their infants were more likely to be aligned when caregivers were more confident about their response (Figure 4). The odds of the two measures aligning increased by 1.8% for every 1% increase in caregivers' self-reported confidence.

## 3.4 | Time Spent With Child

In the final set of preregistered analyses, we tested whether caregivers who spent more time with their child would be more aligned with their child's LWL behavior, both with respect to overall vocabulary and specific word knowledge.

To examine whether time spent between a caregiver and their infant moderates the relation between overall vocabulary and LWL accuracy, we fit a linear model predicting overall accuracy from overall vocabulary size, proportion of time spent together, and their interaction. Infants' overall accuracy was not related to caregiver-reported vocabulary ($p = 0.534$), proportion of time spent together ($p = 0.160$), or the interaction between caregiver-reported vocabulary and proportion of time spent together ($p = 0.404$). These findings suggest that a caregiver's report of vocabulary is not more aligned with overall word recognition accuracy when they spend more time with their child.

We next assessed whether caregiver and LWL behavior were associated when considering the proportion of time spent together for the specific words tested in the LWL task. We fit an LMEM regressing word-level accuracy on caregiver report of word knowledge (−0.5 for does not know the word and 0.5 for understands or says the word), proportion of time spent together, and their interaction, including a by-subject random slope for caregiver report and a by-item random intercept and slope for the interaction between caregiver report and time spent with their child. Infants were more accurate on the LWL task when they had caregivers who reported spending proportionally more time together, controlling for the effect of caregiver report of word knowledge and the interaction effect ($b = 0.204$, $F(1, 118.897) = 4.025$, $p = 0.047$). However, the interaction effect was not a significant predictor of LWL accuracy ($b = -0.242$, $F(1, 18.328) = 1.382$, $p = 0.255$), suggesting that time spent together does not moderate the relation between LWL accuracy and caregiver report of word knowledge. An analogous exploratory model using the alignment score approach as the outcome measure (i.e., 1 if the measures aligned on individual word knowledge and 0 if the measures did not align, as in the final analysis for confidence reported above) revealed that time spent together was not a significant predictor of alignment ($p = 0.755$).

## 3.5 | Exploratory Analyses—LWL Distractors

The preregistered analyses focused on whether caregiver factors influence the alignment between LWL and caregiver-report measures. However, design features of the LWL task may also influence alignment. These factors are poorly understood, and little guidance exists about best practices for study design in order to maximize the validity of LWL measures.

The design of the present study, which included nouns with a range of difficulty (e.g., AoA), provides the opportunity to examine two potentially relevant design features that have not been previously explored. The first is the child's knowledge about the distractor. We reasoned that if the child knows the label for the distractor object but not the target object, they might nevertheless show accurate looking on that trial based on mutual exclusivity (e.g., Markman and Wachtel 1988). The second related issue is mismatches in AoA between targets and distractors. If the target word is well-learned relative to the distractor word, it should be easier for the infant to rule out the distractor as the potential referent, potentially inflating accuracy. However, when the two items are similar in difficulty, gaze behaviormay provide a more veridical measure of word knowledge, potentially increasing alignment with caregivers' reports. In a set of

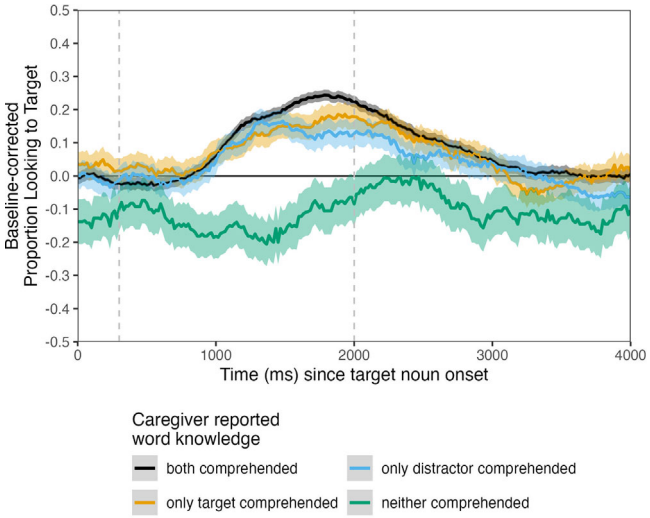


**FIGURE 5** | Baseline-corrected proportion looking to the target unfolding over time. Each line represents gaze behavior as a function of whether infants were reported to know only the target word, only the distractor word, both words, or neither word. Lines were fit using generalized additive models, and error bars reflect standard error of the point estimate. The two dotted lines indicate the critical window of analysis (i.e., 300–2000 ms).

exploratory analyses, we examined whether these two factors influenced alignment.

We investigated the impact of distractor images on alignment by testing whether caregiver-reported knowledge of the distractor word moderated alignment between caregiver-reported knowledge of the target word and LWL accuracy for the target word. To examine this question, we regressed word-level accuracy on caregiver report of target word knowledge (−0.5 coded for "does not know" and 0.5 for "understands and/or says"), caregiver report of the distractor word knowledge (−0.5 coded for "does not know" and 0.5 for "understands and/or says"), and their interaction, including by-subject and by-item random slopes for the interaction term. Caregiver report of target word knowledge was a significant predictor of LWL accuracy when controlling for the effect of distractor word knowledge and their interaction ($b$ = 0.127, $F(1, 1006.95)$ = 11.847, $p < 0.001$). Caregiver report of distractor word knowledge was also a significant predictor of LWL accuracy when controlling for the effect of target word knowledge and their interaction ($b$ = 0.107, $F(1, 470.93)$ = 7.457, $p$ = 0.007). Interestingly, there was a marginally significant interaction effect on LWL accuracy ($b$ = −0.180, $F(1, 19.77)$ = 3.951, $p$ = 0.061), suggesting that the relation between caregiver report of the target word and infant's LWL behavior depended on caregiver report of the distractor word. Infants who were reported as not knowing a target word were more accurate at recognizing the target word when it was paired with a known distractor word (per caregivers' reports) as compared to target words that were paired with an unknown distractor word. Regardless of caregivers' report of the target word, infants were more accurate at recognizing words when caregivers reported that their child understood or said the distractor word (Figure 5).

We also investigated whether a mismatch in AoA between the target word and the distractor word moderated the relation between caregiver report of individual word knowledge and gaze behavior on the LWL. We reasoned that an AoA mismatch may impact the relative ease of rejecting the distractor as the referent for the spoken word. This, in turn, could impact the likelihood that the LWL task provides a valid measure of the child's knowledge about a given target word and result in no relation between caregivers' reports and LWL behavior. Therefore, we predicted that AoA would moderate alignment between the two measures.

To test this hypothesis, we fit an LMEM regressing baseline-corrected proportion of target looking (i.e., accuracy) on caregiver report of word knowledge (−0.5 coded for "does not know" and 0.5 for "understands and/or says"), an AoA difference score (centered), and their interaction, including a by-subject random intercept, a by-subject random slope for the interaction effect, and a by-item random intercept. The AoA difference score was calculated as the difference between the AoA of the target word and the AoA of the distractor word (Range: −11 to 11). AoA was calculated from wordbankr (Braginsky 2024) and defined as the age at which 50% of children are reported to produce the word according to a large database of CDI administrations (Braginsky et al. 2016). Neither caregiver report of word knowledge ($p$ = 0.060) nor AoA difference ($p$ = 0.691) was a statistically significant predictor of LWL accuracy. Interestingly, there was a significant interaction between caregiver report of word knowledge and AoA difference ($b$ = −0.027; $F(1, 104.03)$ = 6.921, $p$ = 0.010), suggesting that the relation between caregiver report of word knowledge and accuracy depended on AoA difference. Specifically, accuracy decreased significantly as AoA difference increased (i.e., the target word has a later AoA than the distractor word), but only for words reported to be understood or said. To put another way, LWL accuracy did not reflect the words caregivers reported as understood or said when the AoA for the target word was much later than the AoA of the distractor word.

## 4 | Discussion

While caregiver-report and LWL tasks are both widely used to measure familiar word knowledge, there is inconclusive evidence as to whether these methods converge on the *specific* words that a child knows (Houston-Price et al. 2007; López Pérez et al. 2025; Lorang et al. 2023; Styles and Plunkett 2009; Syrnyk and Meints 2017; Venker et al. 2016). Each method assesses word knowledge in distinctly different ways. Caregivers' reports draw on accumulated knowledge of their child's experiences with words and referents, and require caregivers to synthesize this rich data into a single discrete response for each word. In contrast, LWL tasks provide a continuous measure of children's ability to associate a word with a particular referent in the context of a single distractor at one moment in time. These differences, among others, likely impact the alignment between these "gold standard" measures of early vocabulary and complicate attempts to reveal early lexical knowledge. In the current study, we investigated specific factors—caregivers' confidence in their responses, the amount of time spent between the caregiver and the infant, and the pairing of referents and distractors in the LWL task—hypothesized to moderate convergent validity between these measures. While caregivers' reports did not predict LWL behavior, neither for overall vocabulary nor for individual words, alignment improved when we included caregivers' confidence in their word-

level reports, as well as infants' knowledge about the distractor words on each LWL trial.

Upon initial review, our findings appear to conflict with mounting evidence that vocabulary size is a key predictor of gaze-behavior during LWL tasks (Fernald et al. 2001, Fernald et al. 2006; Fernald and Marchman 2012; Marchman and Fernald 2008; Weisleder and Fernald 2013; Zangl et al. 2005). However, there are several key design differences to consider that may account for the lack of predictive power of overall vocabulary in the current results. For example, prior studies selected words that were maximally familiar to their participants according to MB-CDI norms (e.g., Fernald et al. 2001). Our design, by contrast, intentionally included words that ranged in AoA to increase the likelihood that we would receive caregiver reports across the word knowledge continuum. This design difference has important implications for the underlying construct being measured. Specifically, prior studies were designed to investigate changes in the speed of language processing across developmental timepoints, while the current study was designed to measure knowledge about specific words at a single timepoint. It is plausible that averaging across known and unknown words masks the facilitative effect of processing skills on later vocabulary size seen in prior studies (e.g., Weisleder and Fernald 2013).

Another important consideration for interpreting the present findings is that the vocabulary percentile for the present sample was lower than average in a sample of typically-developing infants (i.e., 27th percentile). While we believe this is within the range of normal variability for 18- to 20-month-olds, we also recognize that this aspect of the sample may contribute to the lack of association between caregivers' reports and infants' gaze behavior. Indeed, caregivers often underestimate word knowledge in samples of children with lower productive vocabularies (e.g., autism; Lorang et al. 2023; Venker et al. 2016). Thus, the present findings may not extend to participant groups with very large productive vocabularies relative to their age-matched peers.

Our results suggest that gaze behavior and caregiver report of specific words were more aligned when caregivers were more certain about their responses. These data provide initial evidence that certainty reports can improve the validity of caregiver-reported vocabulary, beyond just reliability (as shown in prior work; Hanania 2021; Lorang et al. 2023). Ultimately, these findings suggest that caregivers are valid sources of information about infants' specific word knowledge when they are confident about their responses. Future research is needed to determine what factors influence caregiver confidence. Based on our analyses, confidence does not simply mirror time spent with the child (which did not improve alignment between looking behavior and caregiver report of word knowledge).

It is important to note that our investigation of confidence as a moderator of alignment was limited by relatively high confidence ratings across participants. It is possible that the null effect of confidence for the interaction between confidence and caregivers' reports for overall word knowledge may be due to a lack of variability rather than due to no true effect. Similarly, the limited evidence of an interaction between confidence and caregivers' reports when using a continuous measure of gaze-behavior (rather than alignment score) may be the result of low variability in confidence. One potential consideration for future work is whether different scales of confidence (e.g., Likert vs. 0–100 scale) yield more variability to continue to investigate the effect of confidence on caregiver report of word knowledge.

Examining the pairs of test stimuli in the LWL task also revealed possible explanations for misalignments. In particular, the results indicated that infants successfully recognized target words when they were reported to know the distractor word, even if they were not reported to understand the target word. This suggests that infants' behavior during the LWL task not only reflects their knowledge of the target word, but also reflects their knowledge of the image with which it is paired. Similarly, differences in AoA between target and distractor words influenced the relation between caregiver report of word knowledge and LWL accuracy. Specifically, infants' accuracy decreased as AoA difference increased for words their caregivers reported as understood or said, suggesting that properties of the word pairs tested influenced the task's measurement of word knowledge. Together, these results suggest that a more valid estimate of infants' specific word knowledge may be obtained from an LWL task when the distractor word is unknown or matched in difficulty to the target word. In other words, measurement of an infant's word recognition is more likely to reflect their actual word knowledge when they cannot use their knowledge of the distractor word to rule it out as the referent.

When infants were reported not to know the target word, they were nevertheless able to identify the correct image at above chance levels when they knew the distractor. This behavior suggests that infants were using mutual exclusivity to successfully complete the LWL task. Mutual exclusivity is a well-documented word learning principle of mapping a new label to an unknown object if it is encountered in the presence of a known label-object (Markman and Wachtel 1988). Indeed, a broad literature has demonstrated infants' robust ability to use mutual exclusivity during word-learning tasks with monolingual infants as young as 17 months of age (Grassmann et al. 2015; Horst and Samuelson 2008; Law and Edwards 2015). However, the current findings demonstrate that infants also leverage mutual exclusivity during word recognition tasks. These results challenge an implicit assumption that word recognition solely reflects knowledge of the target word. Rather, in-the-moment word comprehension may be indexing several different underlying processes (e.g., label-object mapping or mutual exclusivity). Future studies examining familiar word knowledge should consider the role of mutual exclusivity when designing LWL tasks as well as when interpreting looking behavior.

Our results provide new insights for understanding the growing literature regarding the assessment of individual word knowledge. Without considering moderators, the current findings could be interpreted as providing additional support for the claim that caregiver-report and LWL measures do not align. This result is not unique to our study; a recent study using archival data from 126 participants conducted by López Pérez and colleagues (2025) demonstrated that although age and overall productive vocabulary size predicted LWL accuracy, caregiver report of the target word did not. Similarly, several studies examining the convergent validity of caregiver report of vocabulary in other languages demonstrate similar findings; caregivers consistently under- or

over-reported word knowledge as compared to LWL behavior (Chai et al. 2023; Ertaş et al. 2024). The current results regarding moderators of alignment provide useful contextualizing information for these prior findings. Without controlling for design factors in each method, prior studies may be capturing the fact that the two measures do indeed assess word knowledge in different ways.

This is the first study, to our knowledge, to provide explicit evidence about *why* caregiver-report and eye-gaze measures frequently fail to converge on knowledge of specific words. That is, assessment of word knowledge can be impacted by design features, such as including measures of caregiver certainty when reporting word knowledge and the specific pairs of test items used on the LWL task. These findings provide insights into how measurement validity can be increased when using these tasks to assess individual word knowledge. With respect to caregivers' reports of particular words, researchers can consider including an additional measure of certainty at the word level and report the extent to which caregivers feel confident about their child's word knowledge. These confidence ratings could also be used to identify a cut-off point at which caregivers' reports are no longer valid measures of specific word knowledge.

In addition, stimuli selection and pairings should be carefully considered when designing LWL tasks to measure word knowledge. LWL tasks are likely to be more valid when target and distractor pairs are matched on AoA and/or the child's existing lexical knowledge. This matching may be accomplished using several different approaches. For example, future studies could explicitly measure word knowledge via caregivers' reports prior to making target and distractor pairs for each child. Alternatively, AoA likely serves as an adequate proxy for whether a child is likely to know a given word, and indeed was a significant moderator of convergent validity in the present study. Yet another option is to select sets of words that are relatively matched on AoA and maximally pair targets and distractors to attempt to wash out any particular pairing effects. Regardless of the specific approach taken, our findings suggest that measuring individual word knowledge can be more valid by accounting for caregiver certainty and LWL stimulus pairings.

In sum, our study provides evidence that when examining infants' knowledge about specific words, caregiver report of word knowledge aligns with eye-gaze behavior when we account for task-relevant factors. Caregivers' judgments converged with their infants' behavior when they were more confident. Other factors that seemed likely to matter, like time spent with the infants, did not influence the convergence between caregiver judgements and infant behavior. Our results also underscore the importance of considering how the design of the LWL task may alter the validity of the method for assessing individual word comprehension. Infants' understanding of both the target and distractor words influenced their ability to fixate the correct referent. Ultimately, our study demonstrates that drawing on insights from both caregiver-report and LWL measures in tandem may provide more reliable and valid estimates of infants' word recognition and early lexical development.


### Acknowledgments

The authors would like to express their gratitude to the families who participated and made this research possible. We would also like to thank Dr. Ron Pomper and Elise Breitfeld for their input on the initial stages of this manuscript.


### Ethics Statement

Ethical approval for this study was obtained from the Minimal Risk Research Institutional Review Board at the University of Wisconsin–Madison. Legal guardians provided written consent prior to the start of the study. Families received compensation for their participation.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The protocol and analyses for this study were preregistered at https://osf.io/c8zws. The stimuli, de-identified data, and analytic scripts that support this research are publicly available on the Open Science Framework (OSF) at https://osf.io/xgnvk/overview?view_only=4c179810233e4066827ade18ba01a15c.

### References


Abelson, R. P., and D. A. Prentice. 1997. "Contrast Tests of Interaction Hypothesis." *Psychological Methods* 2, no. 4: 315–328. https://doi.org/10.1037/1082-989X.2.4.315.

Arunachalam, S., V. Avtushka, R. J. Luyster, and W. Guthrie. 2022. "Consistency and Inconsistency in Caregiver Reporting of Vocabulary." *Language Learning and Development* 18, no. 1: 81–96. https://doi.org/10.1080/15475441.2021.1931233.

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.

Bergelson, E., and D. Swingley. 2012. "At 6-9 Months, Human Infants Know the Meanings of Many Common Nouns." *Proceedings of the National Academy of Sciences of the United States of America* 109, no. 9: 3253–3258. https://doi.org/10.1073/PNAS.1113380109.

Bergelson, E., and D. Swingley. 2015. "Early Word Comprehension in Infants: Replication and Extension." *Language Learning and Development* 11, no. 4: 369–380. https://doi.org/10.1080/15475441.2014.979387.

Brauer, M., and J. J. Curtin. 2018. "Linear Mixed-effects Models and the Analysis of Nonindependent Data: a Unified Framework to Analyze Categorical and Continuous Independent Variables That Vary Within-subjects and/or Within-items." *Psychological Methods* 23, no. 3: 389–411. https://doi.org/10.1037/met0000159.

Braginsky, M., D. Yurovsky, V. A. Marchman, and M. C. Frank. 2016. "From Uh-oh to Tomorrow Predicting Age of Acquisition for Early Words Across Languages." *Proceedings of the Annual Meeting of the Cognitive Science Society* 38. Retrieved from https://escholarship.org/uc/item/9sn4p5f9.

Braginsky, M. 2024. "wordbankr: Accessing the Wordbank Database." R package version 1.0.3.9000. https://langcog.github.io/wordbankr/.

Byers-Heinlein, K., and J. F. Werker. 2009. "Monolingual, Bilingual, Trilingual: Infants' Language Experience Influences the Development of a Word-Learning Heuristic." *Developmental Science* 12, no. 5: 815–823. https://doi.org/10.1111/j.1467-7687.2009.00902.x.

Campbell, J., and D. G. Hall. 2022. "The Scope of Infants' Early Object Word Extensions." *Cognition* 228: 105210. https://doi.org/10.1016/J.COGNITION.2022.105210.

Casey, K., C. E. Potter, C. Lew-Williams, and E. H. Wojcik. 2023. "Moving beyond 'Nouns in the Lab': Using Naturalistic Data to Understand Why Infants' First Words Include Uh-Oh and Hi." *Developmental Psychology* 59, no. 11: 2162–2173. https://doi.org/10.1037/DEV0001630.

Chai, J. H., M. McDonald, and E.-S. Ko. 2023. "Investigating the Convergent Validity of Child Language Assessment Measures: A Korean Study." Preprint, OSF, February 4. https://doi.org/10.31234/OSF.IO/5Q24V.

Clopper, C. J., and E. S. Pearson. 1934. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial." *Biometrika* 26, no. 4: 404. https://doi.org/10.2307/2331986.

Dorai-Raj, S. 2022. "binom: Binomial Confidence Intervals for Several Parameterizations." In *CRAN: Contributed Packages*. Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.binom.

Erdfelder, E., F. Faul, and A. Buchner. 1996. "GPOWER: a General Power Analysis Program." *Behavior Research Methods, Instruments, & Computers* 28: 1–11. https://doi.org/10.3758/BF03203630.

Ertaş, S., A. C. Küntay, and A. Aktan-Erciyes. 2024. "Maternal Underestimations and Overestimations of Their Infants' Word Comprehension: Effects on Mothers' Verbal Input and Infants' Receptive Vocabulary." *Journal of Child Language* : 1–23. https://doi.org/10.1017/S0305000924000576.

Feldman, H. M., T. F. Campbell, M. Kurs-Lasky, et al. 2005. "Concurrent and Predictive Validity of Parent Reports of Child Language at Ages 2 and 3 Years." *Child Development* 76, no. 4: 856–868. https://doi.org/10.1111/J.1467-8624.2005.00882.X.

Fenson, L., V. A. Marchman, D. J. Thal, et al. 2007. *MacArthur-Bates Communicative Development Inventories*. Paul H. Brookes Publishing Co.

Fernald, A., and V. A. Marchman. 2012. "Individual Differences in Lexical Processing at 18 Months Predict Vocabulary Growth in Typically Developing and Late-Talking Toddlers." *Child Development* 83, no. 1: 203–222. https://doi.org/10.1111/J.1467-8624.2011.01692.X.

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. https://doi.org/10.1037/0012-1649.42.1.98.

Fernald, A., D. Swingley, and J. P. Pinto. 2001. "When half a Word Is Enough: Infants Can Recognize Spoken Words Using Partial Phonetic Information." *Child Development* 72, no. 4: 1003–1015. https://doi.org/10.1111/1467-8624.00331.

Fernald, A., R. Zangl, A. L. Portillo, and V. A. Marchman. 2008. "Looking While Listening: Using Eye Movements to Monitor Spoken Language." In *Developmental Psycholinguistics: On-line Methods in Children's Language Processing*, 97–135. John Benjamins Publishing Company. https://www.researchgate.net/publication/267179424.

Frank, M. C., M. Braginksy, D. Yurovsky, and V. A. Marchman. 2021. *Variability and Consistency in Early Language Learning*. MIT Press. https://langcog.github.io/wordbank-book/.

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.

Garrison, H., G. Baudet, E. Breitfeld, A. Aberman, and E. Bergelson. 2020. "Familiarity Plays a Small Role in Noun Comprehension at 12–18 Months." *Infancy* 25, no. 4: 458–477. https://doi.org/10.1111/INFA.12333.

Golinkoff, R. M., W. Ma, L. Song, and K. Hirsh-Pasek. 2013. "Twenty-Five Years Using the Intermodal Preferential Looking Paradigm to Study Language Acquisition: What Have We Learned?" *Perspectives on Psychological Science* 8, no. 3: 316–339. https://doi.org/10.1177/1745691613484936.

Grassmann, S., C. Schulze, and M. Tomasello. 2015. "Children's Level of Word Knowledge Predicts Their Exclusion of Familiar Objects as Referents of Novel Words." *Frontiers in Psychology* 6: 155507. https://doi.org/10.3389/FPSYG.2015.01200/BIBTEX.

Halberda, J. 2003. "The Development of a Word-Learning Strategy." *Cognition* 87, no. 1: B23–B34. https://doi.org/10.1016/S0010-0277(02)00186-5.

Hanania, A. 2021. "Certainty Ratings of Lexical Knowledge on Vocabulary Checklists for Children with Autism Spectrum Disorder." Master's thesis, Michigan State University ProQuest Dissertations & Theses. https://www.proquest.com/docview/2546600393?fromopenview=true&pq-origsite=gscholar&sourcetype=Dissertations%20&%20Theses.

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.

Hendrickson, K., D. Poulin-Dubois, P. Zesiger, and M. Friend. 2017. "Assessing a Continuum of Lexical–Semantic Knowledge in the Second Year of Life: A Multimodal Approach." *Journal of Experimental Child Psychology* 158: 95–111. https://doi.org/10.1016/J.JECP.2017.01.003.

Hirsh-Pasek, K., K. M. Cauley, R. M. Golinkoff, and L. Gordon. 1987. "The Eyes Have It: Lexical and Syntactic Comprehension in a New Paradigm." *Journal of Child Language* 14, no. 1: 23–45. https://doi.org/10.1017/S030500090001271X.

Horst, J. S., and L. K. Samuelson. 2008. "Fast Mapping but Poor Retention by 24-Month-Old Infants." *Infancy* 13, no. 2: 128–157. https://doi.org/10.1080/15250000701795598.

Houston-Price, C., Z. Caloghiris, and E. Raviglione. 2010. "Language Experience Shapes the Development of the Mutual Exclusivity Bias." *Infancy* 15, no. 2: 125–150. https://doi.org/10.1111/J.1532-7078.2009.00009.X.

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.

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: . https://doi.org/10.1098/RSOS.180711.

Kucker, S. C., and M. G. Lorenz. 2025. "Expanding the Footnote: How Minor Methodological Decisions Have Impactful Consequences for Developmental Results." *Frontiers in Developmental Psychology* 3: 1629396. https://doi.org/10.3389/FDPYS.2025.1629396.

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.

López Pérez, M., C. Moore, A. Sander-Montant, and K. Byers-Heinlein. 2025. "Infants' Knowledge of Individual Words: Investigating Links between Parent Report and Looking Time." *Infancy* 30, no. 1: e12641. https://doi.org/10.1111/INFA.12641.

Lorang, E., A. Hanania, and C. E. Venker. 2023. "Parent Certainty and Consistency When Completing Vocabulary Checklists in Young Autistic Children." *Journal of Speech, Language, and Hearing Research* 66, no. 8: 2750–2765. https://doi.org/10.1044/2023_JSLHR-22-00623.

Marchman, V. A., P. S. Dale, and F. Larry. 2023. *MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual (Third)*. Brookes Publishing Co. https://products.brookespublishing.com/MacArthur-Bates-Communicative-Development-Inventories-Users-Guide-and-Technical-Manual-Third-Edition-P1481.aspx.

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.

Markman, E. M. 1990. "Constraints Children Place on Word Meanings." *Cognitive Science* 14, no. 1: 57–77. https://doi.org/10.1207/S15516709COG1401_4.

Markman, E. M., and G. F. Wachtel. 1988. "Children's Use of Mutual Exclusivity to Constrain the Meanings of Words." *Cognitive Psychology* 20, no. 2: 121–157. https://doi.org/10.1016/0010-0285(88)90017-5.

McMurray, B. 2007. "Defusing the Childhood Vocabulary Explosion." *Science* 317, no. 5838: 631. https://doi.org/10.1126/science.1144073.

McRae, K., V. R. De Sa, and M. S. Seidenberg. 1997. "On the Nature and Scope of Featural Representations of Word Meaning." *Journal of Experimental Psychology: General* 126, no. 2: 99–130. https://doi.org/10.1037/0096-3445.126.2.99.

Muradoglu, M., J. R. Cimpian, and A. Cimpian. 2023. "Mixed-Effects Models for Cognitive Development Researchers." *Journal of Cognition and Development* 24, no. 3: 307–340. https://doi.org/10.1080/15248372.2023.2176856.

Nordahl-Hansen, A., A. Kaale, and S. E. Ulvund. 2014. "Language Assessment in Children With Autism Spectrum Disorder: Concurrent Validity Between Report-based Assessments and Direct Tests." *Research in Autism Spectrum Disorders* 8, no. 9: 1100–1106. https://doi.org/10.1016/J.RASD.2014.05.017.

Peters, R., and A. Borovsky. 2019. "Modeling Early Lexico-semantic Network Development: Perceptual Features Matter Most." *Journal of Experimental Psychology: General* 148, no. 4: 763–782. https://doi.org/10.1037/xge0000596.

Pomper, R., and J. R. Saffran. 2019. "Familiar Object Salience Affects Novel Word Learning." *Child Development* 90, no. 2: e246–e262. https://doi.org/10.1111/CDEV.13053.

Potter, C. E., and C. Lew-Williams. 2023. "Frequent vs. Infrequent Words Shape Toddlers' Real-Time Sentence Comprehension." *Journal of Child Language* 51, no. 6: 1478–1488. https://doi.org/10.1017/S0305000923000387.

R Core Team. 2024. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ring, E. D., and L. Fenson. 2000. "The Correspondence Between Parent Report and Child Performance for Receptive and Expressive Vocabulary Beyond Infancy." *First Language* 20, no. 20: 141–159. https://doi.org/10.1177/014272370002005902.

Sachse, S., and W. Von Suchodoletz. 2008. "Early Identification of Language Delay by Direct Language Assessment or Parent Report?" *Journal of Developmental and Behavioral Pediatrics* 29, no. 1: 34–41. https://doi.org/10.1097/DBP.0B013E318146902A.

Smolak, E., K. Hendrickson, P. Zesiger, D. Poulin-Dubois, and M. Friend. 2021. "Visual and Haptic Responses as Measures of Word Comprehension and Speed of Processing in Toddlers: Relative Predictive Utility." *Journal of Experimental Child Psychology* 203: 105032. https://doi.org/10.1016/J.JECP.2020.105032.

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. 2010. "Fast Mapping and Slow Mapping in Children's Word Learning." *Language Learning and Development* 6, no. 3: 179–183. https://doi.org/10.1080/15475441.2010.484412.

Syrnyk, C., and K. Meints. 2017. "Bye-Bye Mummy—Word Comprehension in 9-Month-Old Infants." *British Journal of Developmental Psychology* 35, no. 2: 202–217. https://doi.org/10.1111/BJDP.12157.

Venker, C. E., E. Haebig, J. Edwards, J. R. Saffran, and S. Ellis Weismer. 2016. "Brief Report: Early Lexical Comprehension in Young Children With ASD: Comparing Eye-Gaze Methodology and Parent Report." *Journal of Autism and Developmental Disorders* 46, no. 6: 2260–2266. https://doi.org/10.1007/S10803-016-2747-Z.

Weaver, H., M. Zettersten, and J. R. Saffran. 2024. "Becoming Word Meaning Experts: Infants' Processing of Familiar Words in the Context of Typical and Atypical Exemplars." *Child Development* 95, no. 5: e352–e372. https://doi.org/10.1111/CDEV.14120.

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.

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.

## Supporting Information

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

**Supporting File 01**: desc70088-sup-0001-SuppMat.docx