Received: 23 April 2021 | Revised: 7 April 2022 | Accepted: 9 May 2022

DOI: 10.1111/desc.13290

Developmental Science WILEY

SHORT REPORT

Looking is not enough: Multimodal attention supports the real-time learning of new words

Sara E Schroer | Chen Yu

Department of Psychology, The University of Texas at Austin, Austin, Texas, USA

Correspondence
Sara E Schroer, Department of Psychology, The University of Texas at Austin, Austin, TX, USA.
Email: saraschroer@utexas.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01HD074601, R01HD093792, T32HD007475; National Science Foundation, Grant/Award Number: GRFP DGE-1610403

Abstract

Most research on early language learning focuses on the objects that infants see and the words they hear in their daily lives, although growing evidence suggests that motor development is also closely tied to language development. To study the real-time behaviors required for learning new words during free-flowing toy play, we measured infants' visual attention and manual actions on to-be-learned toys. Parents and 12-to-26-month-old infants wore wireless head-mounted eye trackers, allowing them to move freely around a home-like lab environment. After the play session, infants were tested on their knowledge of object-label mappings. We found that how often parents named objects during play did not predict learning, but instead, it was infants' attention during and around a labeling utterance that predicted whether an object-label mapping was learned. More specifically, we found that infant visual attention alone did not predict word learning. Instead, coordinated, multimodal attention––when infants' hands and eyes were attending to the same object––predicted word learning. Our results implicate a causal pathway through which infants' bodily actions play a critical role in early word learning.

KEYWORDS
attention, eye tracking, multimodal behaviors, parent–infant interaction, word learning

1 | INTRODUCTION

Learning new words seems to be an easy task for infants, but a hard problem for developmental researchers to figure out how infants accomplish the task. Moments when parents use object names to refer to things in the world are often deemed ambiguous and information about word-referent mappings seems fleeting (Trueswell et al., 2016). This narrative begins to shift, however, when we consider the experience of the infant learner. Noisy background information, variability, and referential ambiguity all support learning, rather than hinder it (Bunce & Scott, 2017; Cheung et al., 2021; Twomey et al., 2018). Social cues also help reduce referential ambiguity. When talking about objects in ambiguous contexts, parents use gestural cues (Cheung et al., 2021) and both infant and adult learners can utilize social cues to resolve uncertainty (Baldwin, 1993; MacDonald et al., 2017). The infant's unique view of the world further reduces referential ambiguity.

1.1 | Infants' view of the world supports learning

Recent studies using head-mounted cameras and eye trackers found that the infant's field-of-view has unique properties, differing dramatically from the parent's view at the same moment of toy play: infant's shorter arms result in held objects taking up a large proportion of the infant's field-of-view; those held objects occlude other parts of the infant's visual environment; and manual actions also create more diverse views of objects in the infant's field-of-view than their parent's (e.g., Bambach et al., 2018; Yu & Smith, 2012). The infant's unique view of the world provides the information that is available for learning.

One study using the Human Simulation Paradigm (as in Trueswell et al., 2016) found that ambiguity is reduced and word learning improves when trained from the infant's point-of-view as opposed to a third-person camera view (Yurovsky et al., 2013). Computer vision models similarly learn better from the infant's view than their parent's (e.g., Bambach et al., 2018). The uncertainty at the crux of the word learning "problem" is diminished with social cues, properties of infant's field-of-view, and the infant's own hands.

1.2 | Infants' manual activity creates learning moments

Infant manual activity modulates their learning input, as parents are more likely to label an object when their infants are manipulating it and creating an object-dominant view (Chang & Deák, 2019; Chang et al., 2016; Suanda et al., 2019; West & Iverson, 2017). When parents talk about objects with this infant-generated visual dominance, infants are more likely to learn that object-label mapping (Yu & Smith, 2012). Infant's "real-time" manual activity has a cascading impact on their language development. Fifteen-month-old infants who created more diverse object views had greater vocabulary growth over the next 6 months – but the variability of object views infants saw when their parents were holding the object did not predict language outcomes (Slone et al., 2019). Despite evidence suggesting the importance of hands in shaping informative moments for word learning, no work has directly related infants' visual attention and manual actions to their real-time learning of object-label mappings.

1.3 | Measuring embodied influences on learning

To examine the causal effects of infant visual attention and manual action on word learning, the present study linked infant behaviors during free play to learning outcomes in a test after the play session. Participants played with 10 unfamiliar objects in a home-like laboratory while wearing wireless eye trackers. Using wireless eye trackers granted full mobility to the dyad while still capturing their visual attention in a naturalistically cluttered environment. Infants' knowledge of the object-label mappings was then tested after the play session. This design allowed us to study the types of social and multimodal behaviors that support word learning. One hypothesis is that looking at an object while hearing its name is sufficient for word learning. More embodied hypotheses, however, would predict that holding may also be critical for learning. To test those hypotheses, we measured infant behavior when parents labeled objects and examined whether visual attention, manual action, or the combination of the two was the most predictive of learning. We analyzed infant behavior not just during a labeling utterance, but also before and after the utterance. Analyzing infant behavior before and after labeling could reveal whether parents followed or directed the infant's attention to the labeled object for successful learning, and whether labeling promoted infant's sustained attention to the target object, creating more information for learning.

RESEARCH HIGHLIGHTS

- Wireless head-mounted eye tracking was used to record gaze data from infants and parents during free-flowing play with unfamiliar objects in a home-like lab environment.
- Neither frequency of object labeling nor infant visual attention during and around labeling utterances predicted whether infants learned the object-label mappings.
- Infants' multimodal attention to objects around labeling utterances was the strongest predictor of real-time learning.
- Taking the infant's perspective to study word learning allowed us to find new evidence that suggests a causal pathway through which infants' bodies shape their learning input.

2 | METHODS

2.1 | Participants

Twelve- to-26-month-old infants and parents were recruited from Bloomington, IN, a primarily white, non-Hispanic community of working- and middle-class families in the Midwest of the United States between November 2018 and November 2019. Families were enrolled in a subject database through word-of-mouth and at community events, such as the farmers' market. English did not need to be the participants' primary language and all infant participants were typically developing. Seventy-six percent of recruited infants tolerated wearing the eye tracker. 62% of these infants contributed usable data ($N = 29$; average age = 17.2; 12 F). The remaining infants had eye-tracking data that did not meet quality standards (bad eye image, $N = 8$; unstable eye image, $N = 9$) or did not complete the screen-based test ($N = 1$). Subject information for the 29 infants is provided in Table S1.

Participants were brought into a laboratory decorated to approximate a studio apartment. The HOME Lab (Home-like Observational Multisensory Environment) had three distinct areas in an open floorplan – a colorful play area, a living room, and a kitchenette. Third-person view cameras and microphones were mounted on the walls and ceiling throughout the space. For the current experiment, only the play area (Figure 1a) and an adjoining test room were used. The University Institutional Review Board approved all procedures and parents provided informed consent to participate.

2.2 | Procedure

We asked parents and infants to play with 10 unfamiliar objects for 10 min or until the infant grew fussy (average amount of usable data per participant = 7.12 min [range = 2.22–11.26 min]). We selected

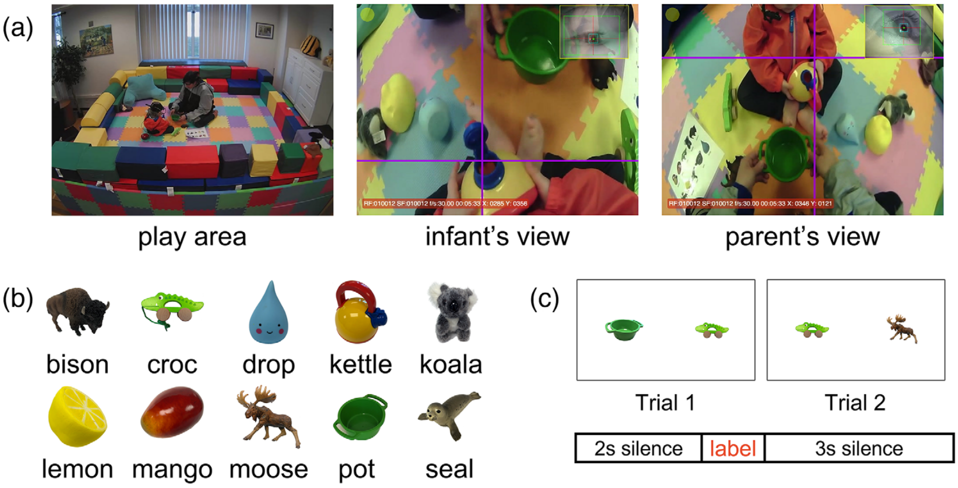

FIGURE 1 (a) Images from three cameras showing the differences in a 3rd-person recording (left), the infant's view (middle), and the parent's view (right) at the same moment. The purple crosshair indicates the location of the participant's gaze. (b) The ten objects and labels. (c) The two trials used to test "croc" in the screen-based task. Trials lasted 7 s: 2 s of silence, then the 1-s labeling utterance, followed by 3-s silence.

everyday objects that infants were unlikely to know the names of as they are not included on the MacArthur–Bates Communicative Development Inventory (MCDI, Fenson et al., 1993; Figure 1b). There were no significant differences in the total amount of time infants spent looking at or holding the 10 toys during the experiment (tested with one-way ANOVAs, *ps* > 0.11). Parents were asked to play as they would at home. To avoid biasing parent behavior during the study, parents were not told in advance that the play session was followed by a word learning test.

While they were playing, dyads wore wireless head-mounted eye trackers (Pupil Labs). Parents wore the "out-of-the-box" eye tracker and infants wore a modified version affixed to a hat. Each eye tracker was attached to an Android smart phone through a USB-C cord. Participants wore custom jackets with a small pocket on the back that the phones were placed into during the experiment. This wire-free set-up allowed participants to move freely throughout the study.

After the experiment, the eye-tracking videos were calibrated (Yarbus software, Positive Science) to produce a crosshair, indicating the location of the participant's in-the-moment gaze in the egocentric view (Figure 1a). To measure visual attention, gaze data were annotated frame-by-frame (30 frames per second) to identify moments the participant was looking at a toy or their social partner's face. The participants' handling of objects was also coded frame-by-frame with an in-house annotation program. Anytime a participant's hand touched an object was coded as manual attention on that object. Parent speech was transcribed at the utterance level using Audacity, following Yu and Smith (2012), from which we identified when parents named an object and analyzed infant visual and manual attention to the named object at those moments.

2.3 | Word learning test

After the play session, infants were tested using a computer monitor and screen-based eye tracker (SMI REDn Scientific Eye Tracker). During each test trial, a target and a distractor would appear on a white screen (Figure 1c). After 2 s of silence, the infants heard the assigned label of the target embedded in the phrase "where's the X?" The labeling utterance lasted approximately 1 s and was followed by 3 s of silence before the trial ended. Infants' knowledge of each object was tested twice, using two different distractors.

A trial was scored if the participant attended to the screen for more than one third of the window after the naming event. Similar to other studies using the looking-while-listening paradigm (e.g., Schwab & Lew-Williams, 2016), trials were scored by dividing the duration of attention to the target object by the total time looking at the two objects during the 3-s window. A trial was considered "correct" if the participant spent a greater proportion of the time looking at the target or "incorrect" if the infant looked more at the distractor during the 3-s window[1]. An object-label mapping was considered "learned" if the infant got both test trials correct and "not learned" if the infant got both trials incorrect for that object. Objects with one correct and one incorrect trial, as well as one or more unusable trials, were excluded from the analyses.

2.4 | Statistical analysis

Mixed-effects logistic regressions were used to assess whether attention to the labeled object during each temporal window (before, during, after) could predict whether an object's label was learned or not learned at test (lmer Test package for R; Kuznetsova et al., 2017). Subjects and objects were included in the models as random effects. All models were compared to the null model (random effect only) using a chi-square test.

3 | RESULTS

Infant participants showed some word learning after the play session. Based on the test, infants, on average, showed learning of 2.3

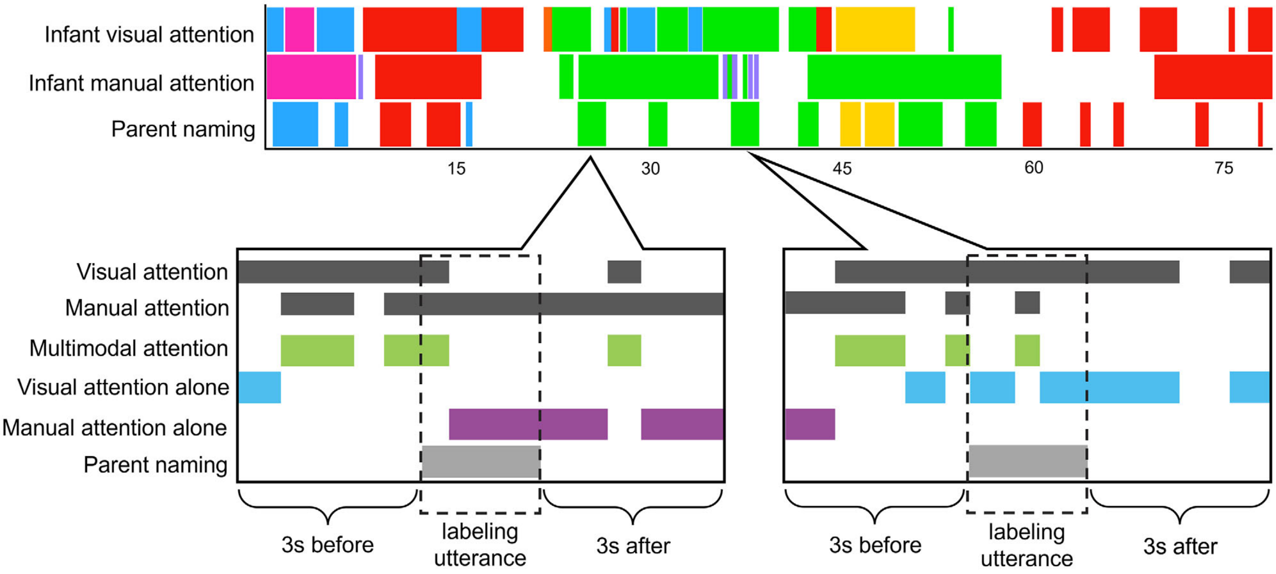

FIGURE 2 The top visualization stream represents 80 s of data from one subject, showing infant behavior and parent naming utterances. Each rectangle represents the onset and offset of a behavior, and the color indicates the object being attended to or talked about. We identified utterances when parents named an object the infant did or did not learn the name of. The two bottom panels show the infant's attention to a target object (dark gray) during the labeling utterance, as well in 3-s windows before and after the utterance. We then identified the moments when the infant was in multimodal attention (green), visual attention alone (blue), or manual attention alone to target object (purple).

(out of 10) words and no learning of 1.9 words. Across all the infants, 66 object instances were categorized as learned and 56 as not learned. The number of objects infants learned and did not learn was not correlated with their age (*ps* > 0.094).

3.1 | Does frequency of parent labeling predict learning?

We first compared how frequently learned and not learned objects were labeled by parents. Across all objects and subjects, there were 256 labeling events of learned objects and 248 labeling events of not learned objects. On average, a learned object was labeled 3.88 times (range: 0–16) and a not learned object was labeled 4.43 times (range: 0–19). There was no significant difference in how many times learned and not learned objects were labeled ($p = 0.479$). Mean length of all naming utterances was 1.25 s. Thus, within the context of this experiment, how many times the infant heard an object's label was not related to whether the mapping was learned. Instead, when infants hear object labels may be more important for learning. To address this question, we analyzed infant attention during and around the labeling events.

3.2 | Does multimodal attention predict learning?

In everyday activities such as toy play, eyes and hands often go together. To examine the potential impacts of multimodal behaviors on word learning, we identified three attention types: multimodal attention (looking and holding at the same time), visual attention alone (looking without holding), and manual attention alone (holding without looking). If just visual attention is sufficient for word learning, then we would expect that *both* multimodal attention and visual attention alone would be significant predictors of learning. Similarly, if just manual attention was sufficient for learning, then both multimodal attention and manual attention alone would predict learning. We calculated the proportion of time infants spent in the three attention types, not just during a labeling utterance, but also within 3 s before and after the utterance. The rationale behind studying the temporal windows separately is to go beyond synchronized behaviors at labeling moments (i.e., infant is attending to the object at the exact moment they hear a label) and examine whether and, if so, how each temporal window may independently contribute to infant word learning (Figure 2).

Before the labeling utterance (Figure 3a and Table 1), infants spent a greater proportion of time in multimodal attention to learned objects ($M_{\text{learned}} = 0.230$, $M_{\text{not learned}} = 0.133$, $p < 0.005$). Conversely, the proportion of time in visual attention alone negatively predicted if the object was learned ($M_{\text{learned}} = 0.131$, $M_{\text{not learned}} = 0.177$, $p = 0.014$). The proportion of time in manual attention alone did not predict learning ($M_{\text{learned}} = 0.217$, $M_{\text{not learned}} = 0.185$, $p = 0.545$). **During the labeling utterance** (Figure 3b), infants spent a greater proportion of time in multimodal attention when the object's label was learned, though this result was only trending on significance ($M_{\text{learned}} = 0.253$, $M_{\text{not learned}} = 0.158$, $p = 0.056$). The proportion of time in visual attention alone ($M_{\text{learned}} = 0.150$, $M_{\text{not learned}} = 0.167$, $p = 0.295$) and manual attention alone ($M_{\text{learned}} = 0.229$, $M_{\text{not learned}} = 0.200$, $p = 0.378$) did not predict learning. **After the labeling utterance** (Figure 3c), multimodal attention still positively predicted whether the object-label mapping was learned ($M_{\text{learned}} = 0.229$, $M_{\text{not learned}} = 0.126$, $p < 0.005$), while visual attention alone ($M_{\text{learned}} = 0.175$, $M_{\text{not learned}} = 0.142$, $p = 0.269$) and manual attention

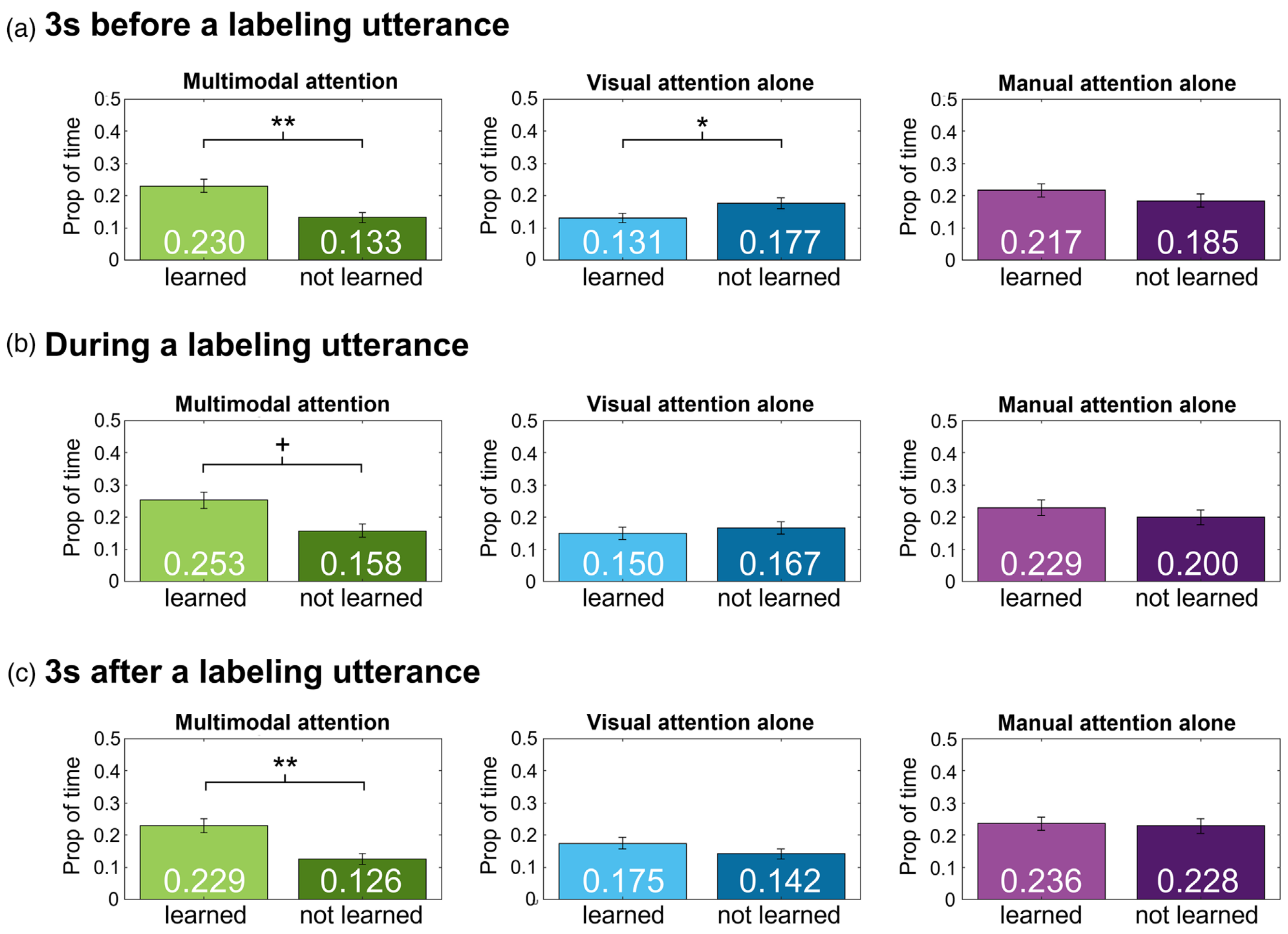

FIGURE 3 The average proportion before (a), during (b), and after the labeling utterance (c) the infant spent attending to the labeled object with multimodal attention (green), visual attention alone (blue), and manual attention alone (purple). The lighter shade is attention to learned objects, and the darker shade is attention to not learned objects. Error bars show standard error. The average value is shown in each bar. + *trending*, **p* < 0.05, ***p* < 0.01

TABLE 1 Summary statistics and regression outputs

	Average proportion of time attending to labeled object		Learned? ~ attention + (1|subject)		
	Learned	Not Learned	β	p	Null model comparison
Multimodal measures of attention					
3 s before labeling utterance					
Multimodal attention	0.230 (*sd* = 0.334)	0.133 (*sd* = 0.240)	1.071	0.004	$\chi^2(1) = 8.478, p = 0.004$
Visual attention alone	0.131 (*sd* = 0.240)	0.177 (*sd* = 0.270)	−1.020	0.014	$\chi^2(1) = 6.050, p = 0.014$
Manual attention alone	0.217 (*sd* = 0.336)	0.185 (*sd* = 0.314)	0.199	0.545	n.s.
During labeling utterance					
Multimodal attention	0.253 (*sd* = 0.402)	0.158 (*sd* = 0.335)	0.547	0.056	$\chi^2(1) = 3.663, p = 0.056$
Visual attention alone	0.150 (*sd* = 0.314)	0.167 (*sd* = 0.301)	−0.352	0.295	n.s.
Manual attention alone	0.229 (*sd* = 0.388)	0.200 (*sd* = 0.368)	0.250	0.378	n.s.
3 s after labeling utterance					
Multimodal attention	0.229 (*sd* = 0.338)	0.126 (*sd* = 0.255)	1.101	0.003	$\chi^2(1) = 9.343, p = 0.002$
Visual attention alone	0.175 (*sd* = 0.284)	0.142 (*sd* = 0.241)	0.446	0.269	n.s.
Manual attention alone	0.236 (*sd* = 0.336)	0.228 (*sd* = 0.357)	0.005	0.987	n.s.

alone ($M_{\text{learned}} = 0.236$, $M_{\text{not learned}} = 0.228$, $p = 0.987$) did not predict learning.

Our results show that multimodal attention, but not visual attention alone nor manual attention alone, was the strongest predictor of word learning. Further, by analyzing each temporal window separately, we found that besides the moments *during* parent labeling, the moments right before and after labeling also matter for learning. When parents follow infant multimodal attention and label the attended object, infants may extend their multimodal attention to the labeled object during and after labeling (see Schroer et al., 2019). By coordinating their visual and manual attention on the same object around naming moments, infants create better opportunities to support real-time learning.

4 | DISCUSSION

It is well-accepted that visually attending to an object while hearing its label is necessary for young learners to build the object-label mapping (e.g., Yu & Smith, 2011). The embodied nature of early word learning has also been suggested through findings such as parents selectively naming objects that infants hold (e.g., Chang et al., 2016; West & Iverson, 2017). However, the present study showed that neither visual attention nor manual attention alone was the best predictor of word learning in the context of toy play. By linking various attention measures during a play session with the results from a learning test immediately after, the presented work identified a causal pathway, suggesting that the perceptual experience of multimodal input created by hand-eye coordination supports the real-time learning of object-label mappings.

4.1 | Infants' hands matter for learning

What are the mechanisms through which multimodal attention supports word learning? Multimodal attention may simply be a stronger indicator of infant overt attention than looking or holding alone. Alternatively, infants may process information in their "hand-space" better than information outside of the hand-space. Research with adults suggests that hands, but not other barriers, scaffold attention by acting as a frame that attracts attention within the hand-space where visual information is processed more efficiently (e.g., Davoli & Brockmole, 2012; Kelly & Brockmole, 2014). Thus, the self-generated multimodal input from hand-eye coordination may reduce distraction in a cluttered visual scene and lead to greater neural representation and increased processing of the object being held (discussed in Davoli & Brockmole, 2012). If the object held by the infant is labeled at the same time, the infant is more likely to build the object-label mapping. What about parents' hands? Although previous studies found that adult gestures and synchronous hand movements improved in-the-moment word learning (e.g., de Villiers Rader & Zukow-Goldring, 2012; Gogate et al., 2000), most experimental studies that emphasized the importance of adult manual actions were designed to not permit the child to touch the objects they were meant to learn. The findings of Brockmole and colleagues may explain why infants' own hands are more important for supporting learning when dyads engage in more naturalistic play (Slone et al., 2019).

4.2 | Manual actions create the learning input

Another mechanism through which multimodal attention supports word learning is that manual actions change the visual input during naming moments. Young children's own bodily actions shape their visual experiences (e.g., Kretch, Franchak, & Adolph, 2014; Yu & Smith, 2012) and create rich opportunities for object exploration. Training studies using the "sticky mittens paradigm" found that providing 3- and 4-month-old infants, who could not yet reach and grasp objects, the opportunity to manipulate objects themselves improved performance on mental rotation tasks and had lasting down-stream effects on object exploration at 15-months-old (Libertus et al., 2016; Slone et al., 2018). Further, manual activities create visual data with more dominant and diverse views of objects that facilitate visual object recognition by computational models (Bambach et al., 2018). In early word learning, manual actions may help infants to identify and segment the named object from a visually cluttered scene, while gazing at that held object at the same time provides high-resolution visual information. Multimodal attention to named objects may thus create a robust pathway from manual action to visual input to successful learning. One way to provide further evidence of this pathway is to analyze the infant's egocentric images and compare the visual properties of the target object during multimodal attention and visual attention alone moments. Furthermore, recent advances in machine learning offer powerful analytics tools to analyze and model visual data collected from the infant's egocentric view (e.g., Orhan et al., 2020; Tsutsui et al., 2020). In a study using the presented data (Amatuni et al., 2021), a model can distinguish frames during learned naming events from frames during not learned naming events. An open question is to what degree this distinction is a result of infants' manual activities. By analyzing visual information in the infant's view, we may discover a visual signature of successful naming moments created by infant's hands, wherein the infant's view of the world is well-suited to word learning.

4.3 | Limitations

The presented work analyzed individual naming instances in parent speech, showing how embodied attention through both eyes and hands facilitates learning object names in naturalistic toy play. Temporal patterns in parent speech, such as repeated naming, can also scaffold learning (Schwab & Lew-Williams, 2016). Future analyses should consider the effect of discourse-level temporal distributions when dyads played with and named objects. Moreover, a few experimental and data analysis decisions may also limit the interpretation of our findings and require follow-up studies in future work. First, to promote a more naturalistic interaction, dyads were given more toys to play with than previous work (e.g., Yu & Smith, 2012)– which may have increased the learning demands on our participants. Second, although we chose target words that are unlikely to be in early vocabulary, it is possible that infants in the study had prior exposure to the 10 words. The effect of prior knowledge can be difficult to predict – prior exposure may be a factor that contributes to the learning outcomes measured at test, but previous work also suggests that infants have worse retention of novel object-label mappings when learning in the context of well-known words (Kucker et al., 2020). Furthermore, despite the wide age range of our participants, neither the infant participant's age nor their concurrent vocabulary size predicted their performance at the test (correlations with MCDI scores reported in Table S2), suggesting that prior knowledge may not have had a direct impact on word learning in the present study. Lastly, we did not collect comprehensive demographic information from our participants, including whether infants were learning English as their first language. Nonetheless, we annotated and examined non-English words used in the play session and found that the three infants that heard any non-English words performed similarly to other subjects at test (these infants still heard the target object labels in English; see Table S1). Future work that considers whether the microlevel behaviors examined in the present study vary across different demographic groups would undoubtedly be a major contribution to the field.

5 | CONCLUSION

We examined multimodal and social factors that support infant word learning in naturalistic parent–infant play when learning is not an exogenous goal of the interaction. Our findings suggest that only studying infant-looking behavior and parent speech is not enough. Considering how infants' bodies shape their visual input will foster a richer, mechanistic understanding of early language acquisition.

ACKNOWLEDGMENTS

This work was supported by NIH R01HD074601 and R01HD093792 to CY. SES was supported by the NSF GRFP (DGE-1610403) and NIH T32HD007475. We thank the Computational Cognition and Learning Lab at Indiana University and the Developmental Intelligence Lab at UT Austin, especially Daniel Pearcy, Hannah Burrell, Dian Zhi, Tian (Linger) Xu, Julia Yurkovic-Harding, Drew Abney, Andrei Amatuni, and Jeremy Borjon for their support in data collection, coding, and many fruitful discussions.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data are not publicly available due to privacy or ethical restrictions, but are available upon request from the corresponding author.

ORCID

Sara E Schroer https://orcid.org/0000-0002-6139-060X

ENDNOTE

[1] To confirm our results, we also used stricter criteria to score trials as correct/incorrect, for example, the infant must look at target for at least 250 ms longer than the distractor to be correct. This increases the number of not learned items and decreases the number of learned items. Nonetheless, the main results of the paper still hold – showing that infants' multimodal attention is the strongest predictor of learning and that visual attention alone in the 3 s before naming negatively predicts learning.

REFERENCES

Amatuni, A., Schroer, S. E., Peters, R. E., Reza, M. A., Zhang, Y., Crandall, D., & Yu, C. (2021). In the-moment visual information from the infant's egocentric view determines the success of infant word learning: A computational study. In Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*(5), 832–843.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2018). Toddler-inspired visual object learning. In *Proceedings of the* Advances in Neural Information Processing Systems (NeurIPS), (pp. 31).

Bunce, J. P., & Scott, R. M. (2017). Finding meaning in a noisy world: Exploring the effects of referential ambiguity and competition on 2.5-year-olds' cross-situational word learning. *Journal of Child Language*, *44*(3), 650–676.

Chang, L., de Barbaro, K., & Deák, G. (2016). Contingencies between infants' gaze, vocal, and manual actions and mothers' object-naming: Longitudinal changes from 4 to 9 months. *Developmental Neuropsychology*, *41*(5–8), 342–361.

Chang, L. M., & Deák, G. O. (2019). Maternal discourse continuity and infants' actions organize 12-month-olds' language exposure during object play. *Developmental Science*, *22*(3), e12770.

Cheung, R. W., Hartley, C., & Monaghan, P. (2021). Caregivers use gesture contingently to support word learning. *Developmental Science*, *24*, e13098.

Davoli, C. C., & Brockmole, J. R. (2012). The hands shield attention from visual interference. *Attention, Perception, & Psychophysics*, *74*(7), 1386–1390.

de Villiers Rader, N., & Zukow-Goldring, P. (2012). Caregivers' gestures direct infant attention during early word learning: The importance of dynamic synchrony. *Language Sciences*, *34*(5), 559–568.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., & Reilly, J. S. (1993). *MacArthur communicative development inventories: User's guide and technical manual*. Paul H. Brookes.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71*(4), 878–894.

Kelly, S. P., & Brockmole, J. R. (2014). Hand proximity differentially affects visual working memory for color and orientation in a binding task. *Frontiers in Psychology*, *5*, 318.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Development*, *85*(4), 1503–1518.

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2020). Sometimes it is better to know less: How known words influence referent selection and retention in 18- to 24-month-old children. *Journal of Experimental Child Psychology*, *189*, 104705.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Libertus, K., Joh, A. S., & Needham, A. W. (2016). Motor training at 3 months affects object exploration 12 months later. *Developmental Science*, *19*(6), 1058–1066.

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology*, *94*, 67–84.

Orhan, A. E., Gupta, V. V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, *33*, 9960–9971.

Schroer, S., Smith, L., & Yu, C. (2019). Examining the multimodal effects of parent speech in parent-infant interactions. In Proceedings of the 41st Annual Meeting of the Cognitive Science Society.

Schwab, J. F., & Lew-Williams, C. (2016). Repetition across successive sentences facilitates young children's word learning. *Developmental Psychology*, *52*(6), 879–886.

Slone, L. K., Moore, D. S., & Johnson, S. P. (2018). Object exploration facilitates 4-month-olds' mental rotation performance. *Plos One*, *13*(8), e0200468.

Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-Generated variability in object images predicts vocabulary growth. *Developmental Science*, *22*, e12816.

Suanda, S. H., Barnhart, M., Smith, L. B., & Yu, C. (2019). The signal in the noise: The visual ecology of parents' object naming. *Infancy*, *24*, 455–476.

Trueswell, J. C., Lin, Y., Armstrong III, B., Cartmill, E. A., Goldin-Meadow, S., & Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent–child interactions. *Cognition*, *148*, 117–135.

Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D., & Yu, C. (2020). A computational model of early word learning from the infant's point of view. In Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.

Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science*, *42*, 413–438.

West, K. L., & Iverson, J. M. (2017). Language learning is hands-on: Exploring links between infants' object manipulation and verbal input. *Cognitive Development*, *43*, 190–200.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*(2), 165–180.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*, 959–966.

SUPPORTING INFORMATION

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

How to cite this article: Schroer, S. E., & Yu, C. (2022). Looking is not enough: Multimodal attention supports the real-time learning of new words. *Developmental Science*, e13290. https://doi.org/10.1111/desc.13290