Beyond Naïve Cue Combination:
Salience and Social Cues in Early Word Learning

Daniel Yurovsky

yurovsky@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank

mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Children learn their first words from social partners, but it is unclear to what extent they are attuned to these partners' social cues. Some theories argue that word learning is fundamentally social from its outset, with even the youngest infants understanding intentions and using them to infer a social partner's target of reference. In contrast, other theories argue that early word learning is largely a perceptual process in which young children map words onto salient objects. One way of unifying these accounts is to model word learning as weighted cue-combination in which children attend to many potential cues to reference, but only gradually learn the correct weight to assign each cue (Hollich, Hirsh-Pasek, & Golinkoff, 2000). We test 3 predictions of a naïve cue-combination account and show each to be incorrect. Thus, while aspects of this unifying account are correct, it must be amended to capture the different dynamics of children's behavior during learning and at test.

Keywords: Language acquisition, word learning, attention, social cues, development

Introduction

How do children learn the meanings of their first words? Infants are situated in a social system from their first day of life, and some theories argue that they leverage this social information from the very outset of word learning (Bloom & Markson, 1998; Waxman & Gelman, 2009). For instance, infants follow direction of gaze by 6-months (D'Entremont, Hains, & Muir, 1997), and are more likely to do so in the presence of other communicative signals (Senju, Csibra, & Johnson, 2008). Further, individual differences in children's gaze-follwoing predict differences in vocabulary development (Brooks & Meltzoff, 2008). In addition, infants appear to representing others beliefs, and these representations affect their expectations by 12-months of age (Vouloumanos, Onishi, & Pogue, 2012). Infants may thus become sensitive to social cues through pre-linguistic experience, and could, in principle, already use these cues from the outset of word learning (Bruner, 1983).

Competing theories argue that early word learning is primarily a perceptual process (Vygotsky, 1978). On these accounts, infants learn words by mapping them onto salient objects in their learning environments (Smith, 2000). Indeed, early child-directed naming events are characterized by multi-modal synchrony: mothers move the objects they label in temporal synchrony with the labels they speak (Gogate, Bahrick, & Watson, 2000), and the degree of synchrony predicts word-object mapping in young infants (Gogate, Bolzani, & Betancourt, 2006).

In an attempting at unification, Golinkoff and Hirsh-Pasek (2006) proposed the *Emergentist Coalition Model*. From this perspective, children are sensitive to a coalition of cues-to-reference: both perceptual cues like visual salience and temporal synchrony, *and* social cues like eye-gaze and pointing. To determine the referent of a speaker's utterance, children *combine* all of the available cues. However, the weight assigned to each cue is not fixed, but changes over the course of development as children learn which cues are the best predictors of referential intent. Early on, children are biased to assign high weight to perceptual cues and low weight to social cues. However, as they learn that cues like social gaze and pointing are better predictors, they gradually assign more weight to these cues and less weight to perceptual cues.

Support for a developmental cue-combination account comes from studies that pit perceptual salience against social information (e.g., a speakers gaze) at different developmental ages. When social gaze conflicts with perceptual salience, 10-month-old infants show no evidence of attending to gaze (Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006). Although 10-month-old infants may be able to follow gaze, they appear to weigh it significantly less than object salience in mapping words to objects. Under similar conditions, 12- and 15-month-olds fail to learn any mappings (Hollich et al., 2000; Houston-Price, Plunkett, & Duffy, 2006). By 19- and 24-months, however, toddlers learn labels for objects cued by gaze even in the presence of salient competitors (Moore, Angelopoulos, & Bennett, 1999; Hollich et al., 2000).

In addition to this empirical support, weighted cue-combination is consistent with properties of our perceptual system (Ernst & Banks, 2002; Jacobs, 2002). Within and across modalities, adults weigh cues in proportion to their predictive power, combining them as predicted by ideal observer models. Using one such model, Frank, Tenenbaum, and Fernald (2013) fit a naïve Bayesian classifier to a corpus of parent-child interactions to determine how perceptual and social cues should be weighed across development.

Weighted cue-combination is an intuitive, computationally simple model of the process of change in early word learning. However, a number of its detailed predictions remain untested. In this paper, we test three:

1. *Developmental change is due to re-weighting across cues*
2. *Cue weights drive attention during learning*
3. *Perceptual cues decrease in weight across development*

Two experiments show that none of these predictions are correct. Thus, while cue-combination captures important insights about early word learning, a naïve version of this account is insufficient.
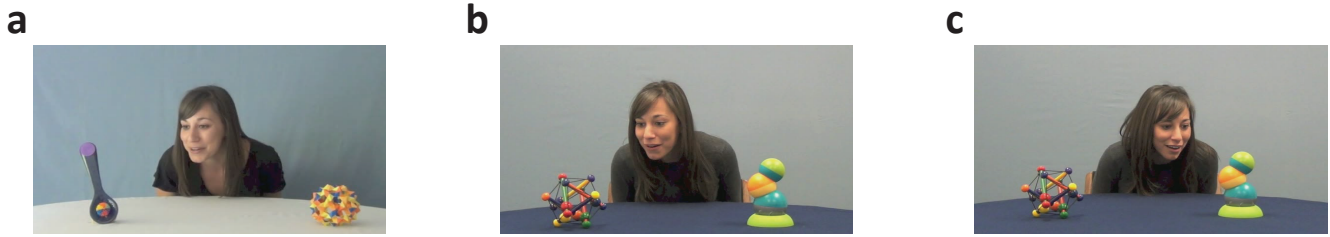
Figure 1: Example learning trials from Experiments 1 and 2. In Experiment 1 (a), the speaker turned towards one of the equally-salient toys and labeled it four times over the course of approximately 10 seconds. In Experiment 2, the speaker produced the same social cues and the same label as in Experiment 1, but the target object was either the more perceptually salient toy (b), or the less perceptually salient toy (c). Across experiments, we were thus able to determine the contributions of both salience and social information to early word learning outcomes.

Experiment 1

In Experiment 1, we set out to measure the development of children's ability to follow and learn from social gaze in the absence of competing salience cues. A naïve cue-combination account, in which developmental changes in cue use result from learning their relative predictive weights, should make a null prediction: children's behavior should not change significantly across development.

Children's eye movements were tracked while they watched a series of naturalistic word-learning videos. In each, children saw a speaker seated at a table between two novel toys. She greeted them, and then turned towards one of the toys and labeled it three times in a short monologue. After these learning trials children were tested for their knowledge of the referent for the new word using the Looking While Listening procedure (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). In addition, similar test trials were administered for known objects to measure children's processing of familiar words.

Method

Stimulus Norming   Thirty-eight adult participants on Amazon Mechanical Turk were presented the toys two at a time and asked to select the one they would rather play with. Each participant made 20 such choices, with toys sampled at random, producing 7.6 responses for each pair of toys. Based on these responses, we selected the two toys that were best balanced against each other (see Figure 1a).

Participants   Parents and their 1–4 year-old children were invited to participate in a short language learning study during their visit to the San Jose Children's Discovery museum. In total, we collected demographic and experimental data from 269 children, 122 of whom were excluded for one or more of the following reasons: abnormal developmental issues ($N = 27$), failure to calibrate ($N = 58$), and less than 75% exposure to English ($N = 36$). The final sample consisted of 27 1–1.5 year olds (9 girls), 19 1.5–2 year olds (7 girls), 38 2–2.5 year olds (13 girls), 26 2.5–3 year olds (10 girls), 15 4–3.5 year olds (9 girls), and 22 3.5–4 year olds (11 girls).

Stimuli   The experiment consisted of two kinds of trials: learning and test. Learning trials were 12s video clips in which a speaker first greeted the the child, and then turned towards one of the two toys on the screen, labeling it three times in a short monologue (Figure 1a). On the first learning trial, for example, the speaker said "Hi there! It's a *modi*. Look at the *modi*. What a nice *modi*."

On each test trial, children saw two objects—one on each side of the screen—and heard a short audio clip of the speaker from the learning trials asking them to find a target object. Each test trial was 7s long, and the target label was heard at 2.75s. On *Familiar* test trials, both the target and distractor were common objects familiar to young children (e.g. book vs. dog). On *Novel* and *Mutual Exclusivity (ME)* test trials, children saw both of the toys from the previous learning trials, and were asked to find either the previously named toy (*modi*), or were asked about a novel label (*dax*). These ME trials were designed as a strong test of mapping formation; looking to the correct target on Novel trials could result from familiarity or preference rather than mapping.

In addition, the experiment contained two calibration checks: short videos in which small dancing stars appeared in four places on the screen. Because eye-tracker calibration can be imprecise, especially with younger children (Morgante, Zolfaghari, & Johnson, 2011), this check allowed us to adjust initial calibration settings to minimize the discrepancy between the behavior children produced and the behavior we analyzed (for details, see Frank, Vul, & Saxe, 2012).

Design and Procedure   The two kinds of trials were designed to measure both how children allocate their attention over the course of learning from a social partner, and what word-object mapping information they extract from these learning events. Consequently, in addition to testing children's ability to correctly fixate the target object on Novel trials, in which they heard the label introduced during learning trials, we measured their fixation behavior when they saw the same objects but were asked to find the referent of a novel word. These Mutual Exclusivity (ME) trials provide a strong measure of children's word learning, because above-chance looking on novel trials could be result of a salience prefer-

ence rather than a mapping between the word and object.

The eye-tracker was first calibrated for each child using a 2-point calibration. Next, children saw four learning trials in which the speaker produced the novel label and disambiguated her target of reference with social gaze. Finally, children saw all of the test trials, in which their knowledge of both familiar and novel word-object mappings was tested. Two calibration checks (described above) were embedded in the learning phase. The entire experiment consisted of 4 learning trials, 8 Familiar, 6 Novel, and 6 ME test trials.

Data Analysis Children's eye movements during both learning and testing were analyzed using a Regions of Interest (ROI) approach. On learning trials, bounding-box ROIs were drawn by a human coder frame-by-frame for the speaker's face and for the two objects. On test trials, a bound-box ROI was drawn for each of the two static images. Children's calibrations were adjusted by fitting a robust linear regression for their fixations during calibration check videos and using this model to transform eye movements during the rest of the experiment (Frank et al., 2012).

Children's learning and test behaviors were quantified by measuring their proportion of looking to each ROI on each trial. To ensure that proportions were representative, individual test trials were excluded from analysis if eye gaze data was missing for more than half of their duration. To compute age-group looking proportions, proportions were computed first for each individual trial, averaged at the individual-child level, and then averaged across children.

Window-of-analysis selection began by coding the point of disambiguation for each trial. This was the onset of the target label for test trials, and the rotation of the speaker's head for learning trials (marked '0' in the graphs in the Results section). The window for each trial began 1s after this point of disambiguation to allow children of all ages enough time to process and continued out to 3s after this point on both learning and test trials. To quantify learning with standard analyses, we aggregated these patterns of looking over time to compute proportion of target looking on each test trial.

Results

In Experiment 1, we address two predictions of naïve cue combination: how cues affect attention during learning, and how weights change across development.

Cue weights drive attention during learning. In line with the cue-combination account, children were attentive to social cues during learning at all ages measured. children were successful at attending to and following the speaker's social gaze even from the youngest ages measured. Children of all ages spent more time looking at the target than at the competitor during learning trials (smallest $t(23) = 3.20$, $p < .01$). However, for all age groups, looks to both target and distractor made up the minority of children's dwell times. Instead, children in all age groups spent more than 50% of their time attending to the speaker's face (Figure 2). Thus, the primary
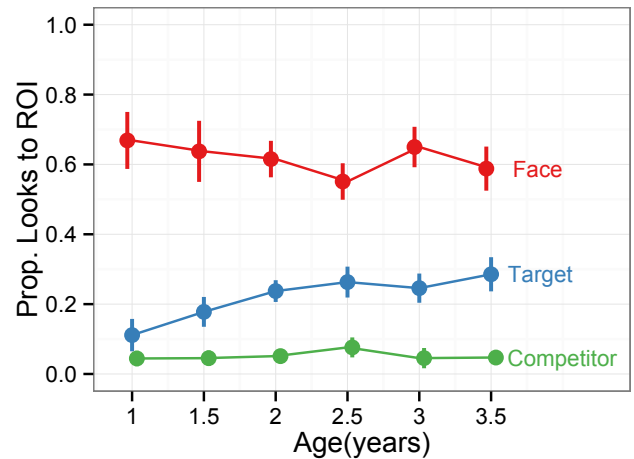

Figure 2: Proportion of children's looking the target toy, competitor toy, and the speaker's face during learning in Experiment 1. Children of all ages spent the majority of the learning trials looking at the speaker's face. Disengaging from the face and fixating the target increased across development. Error bars indicate 95% confidence interval computed by nonparametric bootstrap.

driver of developmental change was not stronger discrimination between the target and distractor, but rather improved ability to disengage from the speaker's face.

Developmental change is due to re-weighting across cues. In line with the naïve cue combination account, attention due to the social cue during learning carried forward to correct mapping at test. Analyses of test trials showed broad success on Familiar, Novel, and ME trials across development. The 1–1.5 year-olds trended towards significance on familiar trials ($t(26) = 1.65$, $p = .11$), and were non-significantly in the correct direction on Novel and ME trials. At all other ages, children looked to the target at above-chance levels on all test trials (smallest $t(17) = 2.10$, $p = .05$).

However, children's abilities both to follow social cues during learning trials and to find the correct target on test trials improved across development. To quantify this improvement, we fit a mixed effects logistic regression to the data (Jaeger, 2008). This analysis revealed significant improvement across age ($\beta = .61$, $z = 4.03$, $p < .001$), as well a significant significant effect of Learning as compared to Novel trials ($\beta = 1.18$, $z = 3.11$, $p < .01$). No other effects or interactions approached significance. Figure 3 shows proportion of looking all kinds of trials at all ages.

Thus, across development, children improved in learning from the social cue, even when it was the only cue available. This suggests that re-weighting across cues is not the only driver of improved word learning.
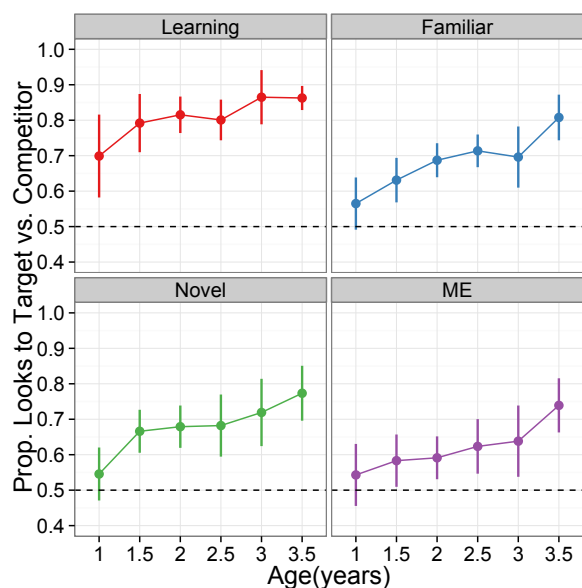
Figure 3: Proportion of time children fixated the correct correct target on each type of test trial in Experiment 1. Children improved on all measures across development. Each dot indicates one half-year age group and each line represents a 95% confidence interval computed by non-parametric bootstrap. A proportion of .5 indicates chance performance.

Discussion

Together, these results provide evidence both of early competence in the use of social gaze to determine the target of a speaker's reference, as well as improvement across development. Further, improvements in gaze-following also paralleled improvements in both finding the referents of these novel words on subsequent learning trials, and also finding the referents of familiar words (Figure 3).

These results thus provide support for one key claim of the developmental cue-combination account: children are sensitive to social cues quite early. Young children could assign small—but non-zero—weight to social cues, and then gradually assign them more credibility over development. However, the results also provide evidence *against* the prediction that cues drive attention, and that developmental change is due to relative re-weighting for two reasons. First, children of all ages found the speaker's face highly engaging, and spent the majority of their time fixating it rather than the referents on learning trials. The primary behavioral development was the ability to disengage from the speaker's face. Second, children showed gradual improvement in fixating the target during both learning and test trials well into their fourth year.

This data could be consistent with a modified version of the cue-combination account in which cues both change in their individual weights and in their relative weights due to learning. However, while children undeniably encounter naming events in their third and fourth years, it would seem unlikely that the process of learning the cue validity of social gaze would extend over such a long period of time.

In Experiment 2 we manipulated the relative salience of the target and distractor objects to which children were exposed. This allowed us to measure how salience affects children's looking during both learning and test, providing a test of all three predictions of the naïve cue-combination account.

Experiment 2

Experiment 2 was identical to Experiment 1 in all respects except for the identity of the novel toys that served as the target and distractor. In contrast to Experiment 1, in which the two toys were balanced in their visual salience, the two toys in Experiment 2 were mismatched. For children in the *Salient* condition, the target was the more interesting toy, and the distractor the less interesting toy. In the *Non-Salient* condition, the identities of the toys were switched—the target was the less salient toy. Thus, Experiment 2 allows us to investigate children's use of social cues to learn new words when they are aligned with salience, and when they are in opposition (as in Hollich et al., 2000; Pruden et al., 2006).

Method

Participants Participants were recruited from the floor of the San Jose Children's Discovery museum as in Experiment 1. This time, however, we focused on the three youngest age groups. In the Salience condition, demographic and experimental data were collected from 117 children, 52 of whom were excluded for one or more of the following reasons: abnormal developmental issues ($N = 13$), failure to calibrate ($N = 25$), less than 75% exposure to English ($N = 33$), and inattentiveness ($N = 2$). The final sample consisted of 22 1-1.5 year olds (11 girls), 21 1.5-2 year olds (10 girls), 19 2-2.5 year olds (9 girls). In the Non-Salience condition, data were collected from 126 children, 71 of whom were excluded for one or more of the following reasons: abnormal developmental issues ($N = 9$), failure to calibrate ($N = 26$), and less than 75% exposure to English ($N = 36$). The final sample consisted of 26 1-1.5 year olds (13 girls), 25 1.5-2 year olds (11 girls), 15 2-2.5 year olds (4 girls).

Stimuli, Design, and Procedure Experimental stimuli were identical to those in Experiment 1, except that the identities of the novel toys were changed and new videos were recorded. The procedure, including the order of the trials, was identical.

Results and Discussion

To determine the effect of perceptual salience on word learning, we compared children's looking in the Salient and Non-Salient conditions not only to each other, but also to the Balanced condition in Experiment 1.

Cue weights drive attention during learning.   In contrast to the prediction of the naïve cue-combination account, children's looking behavior during learning trials was not significantly affected by the salience of the target and distractor. As in Experiment 1, children of all ages spent the more time
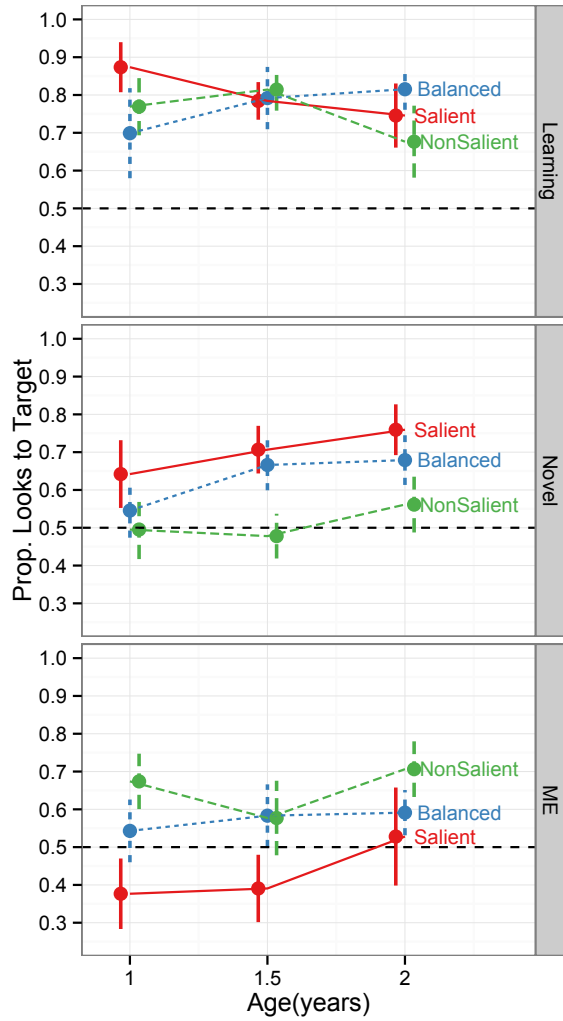
Figure 4: Proportion of time children fixated the correct correct target on Learning and Test trials in Experiments 1 and 2. Salience had effect on looking behavior at test, but relatively little during learning. Each dot indicates one half-year age group and each line represents a 95% confidence interval computed by non-parametric bootstrap. A proportion of .5 indicates chance performance.

looking at the target than the distractor, but looking time to both made up the minority of their dwell time; children spent the majority of learning trials looking at the speaker's face (DOES THIS NEED A FIGURE?).

This null-result could be due to the toys being too similar in their salience, making this a weak test of the cue-combination model. However, salience exerted a strong effect on test trials—children in all age groups were strongly attracted to the salient object. When the target referent was salient, children at all ages looked at it for the majority of the window of analysis on Novel test trials (smallest $t(19) = 2.96$, $p < .01$). When the target was non-salient, no age group look showed evidence of learning on Novel test trials (largest $t(13) = 1.46$, $p = .17$). Mutual-exclusivity (ME) trials showed the opposite pattern. When the target referent was salient, children in the two younger age groups looked at the correct referent on ME trials (the competitor) at *below* chance levels (smallest $t(20) = -2.29$, $p < .05$). In the Non-Salient condition, even the youngest children looked at the correct referent on ME trials at above chance levels (smallest $t(22) = 4.51$, $p < .001$). Figure 4) shows looking behavior in both Experiments 1 and 2 together.

Perceptual cues decrease in weight across development. The effect of perceptual cues at test did not appear to change across development. We fit a mixed-effects logistic regression to the data from all three experiments to determine how how age and experimental condition impacted looking behavior in training and test. After controlling for performance on Familiar trials, this regression showed a significant effect of condition, and an interaction between trial type and condition. Children looked more to the salient object at test regardless of whether it was the target or distractor, and significantly more at the target during learning trials regardless of whether it was salient. Further, there was no interaction of any of these factors with age (Table 1).

Developmental change is due to re-weighting across cues. Together with the t-tests above, this analysis suggests that children are not learning to relative weights on salience and social cues over the course of development. While salience certainly plays a role in directing looking behavior, it does not appear to play a role during learning itself. Instead, salience appears to have a strong effect during test. In the absence of any social information, salience directs children's attention in a way that does not appear to change over early development.

Table 1: Mixed-effects Regression Coefficients Predicting Looking Behavior in Experiments 1 and 2.

Predictor	Value (SE)	t-value	Sig.	
Intercept	-.65 (.63)	-1.04	$p = .3$	
Age (yrs)	.42 (.27)	1.58	$p = .11$	
Familiar	1.55 (.73)	2.13	$p < .05$	*
Salient	.96 (.48)	1.98	$p < .05$	*
NonSalient	-.97 (.37)	-2.63	$p < .01$	**
Learning	1.55 (.73)	2.13	$p < .05$	*
ME	-.29 (.36)	-.80	$p = .42$	
Salient*Learn	-.03 (.84)	-.035	$p = .97$	
NonSal*Learn	1.09 (.65)	-1.65	$p = .85$	
Sal*ME	-2.28 (.61)	-3.73	$p = -.09$	.
NonSal*ME	1.67 (.54)	3.07	$p < .01$	**

Conclusion

Main ideas: face is relevant, attended to, and used to find the target. The reason why could change over development (e.g. face itself could be salient), but either way "social cues" get used.

Salience definitely matters, but mostly at test. Could be a cue-weighting model in which once the face is gone the salience cue gets high weight. But something about cue weighting seems wrong based on learning trial behavior. Unclear how to predict developmental differences unless cues are not normalized. Also, in general, this kind of model really misses the inherent temporal aspects of word learning.

Salience could still be used in the absence of social information for either smart or dumb reasons, and might end up being adaptive. But it seems pretty clear that we're not seeing a re-weighting across development of these two cues.

Seems like what we really want to pay attention to are memory, attentional control, and timing. Probably all of these are developing?

Maybe the two timescales idea (Frank, Goodman, & Tenenbaum, 2009; McMurray, Horst, & Samuelson, 2012; Yu & Smith, 2012) is the right way to think about this kind of thing. Performing above competence?

Acknowledgments

References

Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences*, *2*, 67–73.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. *Journal of Child Language*, *35*, 207–220.

Bruner, J. (1983). *Childs talk: Learning to use language*. Oxford: Oxford University Press.

D'Entremont, B., Hains, S. M. J., & Muir, D. W. (1997). A Demonstration of Gaze Following in 3- to 6-Month-Olds. *Infant Behavior and Development*, *20*, 569–572.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*, 228-231.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Frank, M. C., Vul, E., & Saxe, R. (2012). Measuring the development of social attention using free-viewing. *Infancy*, *17*, 355–375.

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*, 878–94.

Gogate, L. J., Bolzani, L. H., & Betancourt, E. A. (2006). Attention to Maternal Multimodal Naming by 6- to 8-Month-Old Infants and Learning of Word-Object Relations. *Infancy*, *9*, 259–289.

Golinkoff, R. M., & Hirsh-Pasek, K. (2006). Baby Wordsmith: From Associationist to Social Sophisticate. *Psychological Science*, *15*, 30–34.

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning. *Monographs of the Society of Research in Child Development*.

Houston-Price, C., Plunkett, K., & Duffy, H. (2006). The use of social and salience cues in early word learning. *Journal of experimental child psychology*, *95*, 27–55.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*(8), 345–350.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831–877.

Moore, C., Angelopoulos, M., & Bennett, P. (1999). Word learning in the context of referential and salience cues. *Developmental psychology*, *35*, 60–8.

Morgante, J. D., Zolfaghari, R., & Johnson, S. P. (2011). A Critical Test of Temporal and Spatial Accuracy of the Tobii T60XL Eye Tracker. *Infancy*, 1–24.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development*, *77*, 266–280.

Senju, A., Csibra, G., & Johnson, M. H. (2008). Understanding the referential nature of looking: infants' preference for object-directed gaze. *Cognition*, *108*, 303–19.

Smith, L. B. (2000). How to learn words: An associative crane. In R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the Word Learning Barrier* (pp. 51–80). Oxford, UK: Oxford University Press.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, *109*(32), 12933–12937.

Vygotsky, L. (1978). *Mind and society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Science*, *13*, 258–263.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, *119*, 21–39.