Combining Cues in Novel Word Learning

By

Ron Pomper

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2020

Date of final oral examination: 06/25/2020

The dissertation is approved by the following members of the Final Oral Committee:

Jenny Saffran, Professor, Psychology

Martha Alibali, Professor, Psychology

Margarita Kaushanskaya, Professor, Communication Sciences Disorders

Karl Rosengren, Professor, Psychology

Table of Contents

Acknowledgements

First and foremost, I would like to thank the families and children who participated in this research. Without your support, none of this work would be possible. Thank you for making the time to come play silly (and sometimes boring) games in the lab.

I am tremendously grateful for everyone who has been a part of the Infant Learning Lab past and present. Thanks to Erin Long, Rachel Reynders, Sara Oakley, Jing Shen, and many, many others for their help in recruiting participants, recording stimuli, and collecting data. I have benefited tremendously from being able to work and grow alongside my fellow lab mates – Martin Zettersten, Viri Benitez, Lynn Perry, Erica Wocjik, Tianlin Wang, Chris Potter, Brianna McMillan, Desia Bacon, and Haley Weaver. Thank you for the endless conversations about research, technical support, laughter, and the wonderful community we created.

I would also like to thank members of my committee Martha Alibali, Karl Rosengren, and Margarita Kuashanskaya. This research is vastly better because of your feedback, insights, and suggestions over the years. More importantly, thank you all for your mentorship. For many years, you welcomed me into your classrooms, lab meetings, lab spaces, and homes. I have learned and continue to learn so much from each of you.

Most importantly, I would like to thank my advisor, Jenny Saffran. Thank you for helping me grow as a researcher, teacher, academic, and human being. You have the incredible ability to provide the right balance of support and space for independence and have kept me in a Zone of Proximal Development for 7 years now. I could not have asked for a better mentor and am so grateful that I decided, so long ago, to come to Madison to work with you.

Lastly, I would like to thank my family. Thank you, mom and dad, for instilling in me the curiosity, stubbornness, and perfectionist tendencies that are perfect for a career in research.

Your unconditional support has helped me through the lowest lows and highest highs of the roller coaster ride that is graduate school. Finally, Phoebe, thank you for your love, patience, and kindness. You have shown me how to be more reflective, intentional, and kinder to myself. You have helped me to thrive, rather than survive, through graduate school.

Abstract

In order to learn new words, children must successfully identify the referents of the words. This is not a trivial task, because when children hear a new word, they are surrounded by many potential referents. Decades of research has examined different cues that children use to solve this problem. The majority of this research has either tested individual cues in isolation or put cues in conflict. The current studies examine how cooperating cues affect word learning. In both studies, three-year-old children watched videos of an adult labeling either a familiar object with a known name or a novel object with a novel name from a collection of objects on a table. In each video different combinations of cues were available to help children identify the referents of the novel words. On subsequent trials children were shown images of the novel objects and heard sentences identifying one by name. Children's eye movements were recorded during both phases to measure their accuracy in fixating the target object. In Study 1, we found that the presence of a second cue improved children's accuracy in fixating the referent during teaching videos, even though individual cues were sufficient. Children's accuracy on test trials, however, was unaffected by the number of cues that were present during teaching. In Study 2, individual cues were not sufficient – children could only identify the referent during teaching videos by using multiple cues. Children's accuracy in fixating the target object on subsequent test trials was significantly higher than chance. Together, these results indicate that the effect of multiple, cooperating cues on word learning is varied. When a single cue will suffice, the addition of a second cue is redundant and does not affect word learning. When single cues are insufficient, however, children can combine multiple, imperfect cues in order to learn the names of novel objects.

Introduction

Children learn many words through indirect exposure, rather than didactic learning. To do so, however, children must identify the referents of new words on their own. This is not a trivial task, because when children hear a new word they are often surrounded by many potential referents. A large body of research has examined how children use different sources of information in their environment to solve this problem of referent selection. There are many different cues that children leverage to identify the correct referent of a novel word. The majority of these cues can be broadly organized into three different accounts: Constraint-Based, Social-Pragmatic, and Associationist (Hollich et al., 2000). These accounts of word learning are polarizing in that they emphasize one type of cue to the relative exclusion of others. This is in part based on the belief that other cues are insufficient (e.g., a speaker's gaze can help identify the referent, but cannot resolve whether the word labels the entire object, a part of the object, or an attribute), as well as the belief that the importance of other cues is overstated (e.g., constraints are necessary in impoverished experiments, but not in socially rich naturalistic word learning, Tomasello, 2000). A brief overview of each account is provided, before further examining the specific cues that are relevant to the current studies.

Cues in Isolation

Proponents of the Constraint-Based approach believe that children do not consider *all* possible referents of a novel word, because of assumptions or constraints that eliminate certain hypotheses (Clark, 1983; Markman, 1987). For instance, children assume that a novel word labels a whole object, rather than part of the object or an attribute of the object (Markman & Wachtel, 1988; Soja, Carey & Spelke, 1991). Furthermore, children assume that a novel word

labels a taxonomic category, rather than a thematic category, and that the taxonomic category is at the object level (e.g., dog) rather than the superordinate (e.g., animal) or subordinate levels (e.g., Lassie; Golinkoff, Shuff-Bailey, Olguin & Ruan, 1995; Markman & Hutchinson, 1984; Waxman & Gelman, 1986). Additionally, children assume that a novel word labels a novel object, rather than a familiar object with a known name (Markman & Wachtel, 1988). This last assumption has been labeled using various terms but is frequently referred to as Mutual Exclusivity (ME) and will be described in greater detail in subsequent sections. By removing many possibilities and narrowing the hypothesis space, ME and other constraints can help to facilitate word learning.

Proponents of the Social-Pragmatic approach emphasize that children do not learn words on their own, but rather through interactions with language experts (i.e., adults). These experts provide social cues (e.g., by looking, pointing, or holding the referent they are labeling) that help infants and children identify the correct referents of novel words. Children use these cues to correctly identify the referent of a novel word. Moreover, they will only learn words when a speaker intentionally labels, but not accidentally labels, a novel action or object (Tomasello & Barton, 1994). Children's ability to use a speaker's gaze in order to learn new words will be examined in greater detail in subsequent sections (Baldwin, 1991; 1993). Social-pragmatists believe that the Constraint-Based account over-exaggerates the amount of ambiguity in word learning; of the infinite number of hypothetical meanings, most are not plausible or even rationale in a given social interaction.

Finally, proponents of the Associationist approach suggest that children rely on domain-general skills like attention and memory to learn new words (Bloom, 2000; Samuelson & Smith, 2000). Social cues (like gaze) and linguistic constraints (like ME) work because they make the

target novel object more salient, but this attentional highlighting can be achieved by other means (e.g., by placing the target novel object on a more visually salient tablecloth; Samuelson & Smith, 1998). Moreover, children may *not* be able to unambiguously select the referent after a single exposure to a novel word but can after multiple exposures if the word consistently co-occurs more often with its referent than with other objects in the environment. Infants track these word-object co-occurrences to correctly identify the referent of new words (Smith & Yu, 2008). Constraint-Based and Social Pragmatic accounts of word learning may not be necessary, because children are able to resolve any referential ambiguity over time. Word learning can therefore be a gradual process, rather than an all-or-nothing opportunity that requires children to use cues like the speaker's gaze or ME to immediately determine the referent of a novel word.

Children's success in referent selection, whether after a single exposure or many, does *not* provide evidence of word learning. Successfully identifying the referent of a novel word does not ensure that children will learn and retain the word-object mapping. To actually demonstrate word learning, children must be able to subsequently identify the novel object *in the absence* of any cues. This distinction is not trivial – past research has found that children's success in referent selection does not guarantee success in word learning (e.g., Bion, Borovsky & Fernald, 2013; Hollich et al. 2000; Horst & Samuelson, 2008). While some research has assessed both referent selection and word learning, other research has just tested referent selection. This distinction is carefully noted in descriptions of prior research in subsequent sections.

While there are many cues that children use to aid word learning, the current studies focus on just two: ME and speaker gaze. These cues come from different accounts – Constraint-Based (ME) and Social-Pragmatic (gaze) – and there is a rich literature of research involving each of these cues, both separately and together. Moreover, these two cues are compatible in that

they allow children to correctly identify a referent following a single exposure – what's been termed fast mapping (Carey & Bartlett, 1978). There are, however, many other cues that affect word learning. Some of these cues have been shown to interact to affect adults' word learning (e.g., MacDonald, Yurovsky & Frank, 2017). Therefore, the current research represents an important step, but only a first step, in examining how cooperating cues affect children's word learning. We will return to this issue in the General Discussion.

Mutual Exclusivity

When shown a novel object and a familiar object with a known name, children and adults select the novel object as the referent of a novel word. This was first shown by Vincent-Smith, Bricker and Bricker (1974), but is most strongly associated with Markman and Wachtel (1988). Markman and Wachtel proposed that this behavior was the product of a word-learning constraint that helps narrow down the hypothesis space. The constraint is that children assume labels are mutually exclusive (ME) – each object only has one label. Since the familiar object already has a name, the novel word must refer to the novel object without a name. This proposal was not without controversy. Indeed, ME cannot always be true, because the same object will have multiple labels, just at different levels of meaning (e.g., Lassie is also a Rough Collie, a dog, and an animal).

Since Markman and Wachtel (1988), many psychologists have proposed and tested different mechanisms that may underlie the tendency to select the novel object as the referent of the novel word. This tendency could be the result of a word-learning constraint that is either a default assumption (Markman, 1992) or something that's learned from experience (Mervis & Bertrand, 1994). Rather than a constraint, however, this tendency could instead reflect social-

pragmatic reasoning: if the adult wanted the familiar object with the known name, then they would have used the known name (Diesendruck & Markson, 2001). Yet another alternative is that this tendency can be explained by domain-general mechanisms like attention (i.e., an Associationist account). If children are biased to attend to novel things, then they may select the novel object without even considering/rejecting the familiar object (Golinkoff, Mervis & Hirsh-Pasek, 1994; Mervis & Bertrand, 1994; Mather & Plunkett, 2010; Horst, Samuelson, Kucker & McMurray, 2011). As previously mentioned, many different terms have been used to describe children's tendency to select novel objects as the referents of novel words; each term is usually associated with a different proposed mechanism. For the sake of clarity, I will use the term ME; I use the term agnostically, however, because the underlying mechanism is irrelevant to the current studies.

In addition to the debate surrounding the proposed mechanisms underlying ME, the age at which ME first emerges has been debated. Across different experiments, there has been substantial variability in the youngest age at which infants will demonstrate ME. In some experiments, infants as young as 10 months of age succeed in ME (Mather & Plunkett, 2010), while in other experiments 14- and 16-month-olds (Halberda, 2003), 18-month-olds (Bion, Borovsky, and Fernald, 2013) and even 24-month-olds (Merriman & Bowman, 1989) *fail* in ME. These differences may be due methodological differences in procedures (e.g., whether children were or were not pre-familiarized with the novel objects) and task demands (Woodward & Markman, 1991). For instance, 2-year-old children succeed in using ME to identify the correct referents of novel words when there are fewer novel words and objects (Golinkoff et al. 1985; Hutchinson, 1986), but not when there are more novel words and objects in the experiment (Merriman & Bowman, 1989). One consistent finding, however, is that success in ME improves

with age (Merriman & Bowman, 1989; Markman & Wachtel, 1988). These age-related improvements are likely driven by improvements in children's vocabulary. Within an age group, children with larger vocabularies are more successful in ME than their peers with smaller vocabularies (Law & Edwards, 2015). More specifically, improvements in ME may be driven by increases in the strength of children's knowledge of the familiar objects' labels. Children are more successful in ME when the label for the familiar object is in both their expressive and receptive vocabulary, compared to just their receptive vocabulary; moreover, this knowledge of the familiar word's label is a stronger predictor of ME performance than age (Grassmann, Schulze & Tomasello, 2015).

Subsequent research on ME has moved beyond the mechanistic and age-related debates towards examining the utility of ME in word learning. Children can use ME to identify the referents of novel words, but this does this guarantee learning. The majority of the early research on ME cannot address this, because children were not subsequently tested on their ability to identify the novel object using its name in the *absence* of ME (i.e., when presented with another novel object). Experiments that have included subsequent tests of retention have found that although 24-month-old children succeed in referent selection, they fail when their retention is tested after even a brief delay (Horst & Samuelson, 2008). By 30 months of age, however, children succeed in retaining the word-object mapping over a short delay (Bion, Borovsky & Fernald, 2013; Horst, Scott & Pollard, 2010).

Finally, more recent work has examined how changing different properties of the familiar objects affects children's success in ME and word learning. Recall that at 24 months of age, children succeed in ME, yet fail in retention (Horst & Samuelson, 2008). 24-month-old children succeed in retaining novel word-object pairings, however, if the novel object is illuminated by a

light after ME (Axelsson, Churchley & Horst, 2012). Changing the number of familiar objects that are present does not affect children's success in ME, but does affect word learning (Horst, Scott & Pollard, 2010). While 30-month-olds succeed in learning and retaining word-object mappings when ME involves two familiar objects with known names, they fail in retaining word-objecting mappings when ME involves three or four familiar objects. Similarly, repetition in the use of familiar objects does not affect children's success in ME, but does affect word learning (Axelsson & Horst, 2014). While 36-month-olds are successful in remembering the names of novel objects when familiar objects are repeated during ME (i.e., if a novel object is always paired with the same familiar object), they fail in retaining word-object mappings when different familiar objects are used ME (i.e., if a novel object is always paired with different familiar objects each time). Finally, manipulating the salience of the familiar objects during ME affects children's success in both ME and word learning. 40-month-olds are less accurate in identifying the referent of a novel word when the familiar object is more salient compared to when the familiar object is less salient (Pomper & Saffran, 2018). While 40-month-olds are successful in remembering the names of novel objects when the familiar objects were less salient during ME, they fail in retaining the word-object mappings when familiar objects were more salient during ME.

Taken together, these experiments demonstrate that children's ability to identify the referent of novel words using ME improves with age, that success in referent selection does not guarantee learning, and that manipulating the familiar objects may not affect children's success in referent selection, but does affect learning. Children's ability to identify the referent of a novel word is necessary, but not always sufficient to ensure word learning.

Speaker Gaze

When children first hear a novel word, they are not always surrounded by familiar objects; instead, there may be several novel objects nearby. In these situations, however, children can use a speaker's gaze to identify which of the many novel objects is the intended referent of a novel word. This ability develops over an extended period of time. At 10 months of age, infants do not use a speaker's gaze, but rather differences in visual salience to identify the referent of a novel word – mapping a novel word to the more visually salient object in a pair, regardless of which object the speaker was looking at (Pruden, Hirsh-Pasek, Golinkoff & Hennon, 2006). By 12 months of age, infants are able to identify the referent of a novel word if the speaker both *looks* and *points* at the novel object, but not when the speaker only looks at the novel object (Hollich et al., 2000). At this age, children are not able to learn and retain the word-object mapping if the speaker *points* to the object but can if the speaker *holds* the object. By 14 months of age, infants successfully learn and retain word-object mappings when the speaker looks at and labels an object that the child is holding (Baldwin, 1993a). 19-month-olds can use a speaker's gaze without temporal contiguity (Akhtar & Tomsello, 1996; Baldwin, 1993b) and 24-month-olds can use a speaker's gaze to successfully learn and remember novel words even when the target novel object is less salient than the distractor novel object (Hollich et al., 2000; Moore, Angelopolous & Bennett, 1999). 30-month-olds use a speaker's gaze to learn new words even when they are overhearing a conversation between the speaker and another adult (Akhtar, Jipson & Callanan, 2001). Finally, by 36 months of age, toddlers are able to use a speaker's gaze to learn new words even when that gaze is subtle and only briefly available (Yurovsky, Wade & Frank, 2013).

As with ME, these results reveal a dissociation between infants' and children's ability to identify the referent of a speaker's gaze and their ability to learn and retain the word-referent mapping. Moreover, these findings reveal a developmental pattern in which young infants need multiple social cues in order to learn new words. With age and more experience, however, older children are able to use *only* a speaker's gaze to successfully learn new words. Social cues like pointing and holding may be less frequent, but they often come packaged with speaker gaze and their presence provides a more reliable indication of referential intent. Speaker gaze is not a spatially precise cue – it is often difficult to precisely identify the target of a person's gaze (Frank, Tenenbaum & Fernald, 2013). Indeed, 12-month-olds coordinate attention by attending to a speaker's hands rather than their gaze (Yu & Smith, 2013).

The developmental changes in infants' ability to use speaker gaze to learn new words are likely connected to developmental changes in social cognition (Tomasello, 1999). By 9 months of age, infants understand that an adult's reaching behavior is goal-directed (Woodward, 1998). It's not until 12 months of age, however, that infants understand that an adult's gaze is goal-directed (Woodward, 2003). With increases in age, children also develop a greater conceptual understanding of others' gaze. For instance, 9-month-olds follow an adult's gaze regardless of whether the adult's eyes are open or closed, while 10-month-olds follow an adult's gaze only when the adult's eyes are open (Brooks & Meltzoff, 2005). Similarly, 12-month-olds will follow an adult's head turn regardless of whether the adult is wearing a blindfold or a headband, while 14-month-olds will only follow a head turn when the adult is wearing a headband (Brooks & Meltzoff, 2002). In addition to these changes in infants' understanding of gaze, there are also developmental changes in infants' ability to establish joint attention. Infants progress from being able to share joint attention when an adult follows in on their focus of attention (between 9 and

11 months of age), to following adults' points and gaze to establish joint attention (between 11 and 12 months of age), and finally directing adults' attention by pointing to establish joint attention (between 12 and 13 months of age; Carpenter, Nagell, Tomasello, Butterworth & Moore, 1998).

Caregivers are sensitive (implicitly or otherwise) to these developmental trajectories when they interact with infants. When interacting with 6- to 10-month-olds, the majority of mothers' utterances are in response to something done by the child (Harris, Jones & Grant, 1983). Moreover, parents naturally use touching and manipulation when teaching their children new words. In one observational study, mothers simultaneously manipulated the referent of their speech 73 to 95% of the time (Messer, 1978). This simultaneous manipulation has been called "multimodal motherese" and occurs more often when mothers are using novel words compared to familiar words that their infants already know (Gogate, Bahrick & Watson, 2000; Masur, 1997). Moreover, mothers use multimodal motherese significantly more often for younger (5- to 8-month-olds, ~ 76%) compared to older infants (9- to 17-month-olds, ~ 36%; Gogate, Bahrick & Watson, 2000). Infant's contributions to interactions also increase with age – mothers name objects that are held by their child significantly more often when their child is older (21- to 30-month-olds) compared to younger (5- to 18-month-olds; Gogate, Bahrick & Watson, 2000). Thus, as infants become more competent in following and then directing adults' attention, caregivers adjust, providing fewer social cues and following their child's lead.

Correlational evidence suggests that these social cognitive abilities are important for infants' and children's language development. Individual differences in infants' understanding of social cues and their caregivers' sensitivity to these changes are related to infants' subsequent language development. Infants who are better able to follow an adult's gaze at 10 to 11 months

of age have subsequently larger vocabulary growth (Brooks & Meltzoff, 2008). Similarly, infants whose mothers were more responsive to their vocalizations between 9 to 13 months of age have subsequently larger vocabulary growth (Tamis-LeMonda, Bornstein & Baumwell, 2001). These individual differences account for a large amount of variance in children's language outcomes. In one instance, individual differences in the amount of joint attention between children and their caregivers accounted for nearly 60% of the variance in children's subsequent vocabulary growth (Carpenter, Nagell, Tomasello, Butterworth & Moore, 1998).

Word learning does not occur in a vacuum, rather children are learning words from adults. When both children and adults are jointly attending to the same object, there is little to no referential ambiguity. Taken together, this research reveals that both children and adults provide social cues like gaze, pointing, and holding to establishing joint attention. Children's ability to identify the referents of novel words using more subtle social cues like gaze improves with age. Success in identifying the focus of a speaker's gaze, however, does not ensure learning. This pattern of results is strikingly similar in many ways to the research involving ME.

Cues in Conflict

In most of the research on word learning, including the research summarized above, individual word learning cues have been tested in isolation. This is important, because it allows researchers to determine whether a specific cue is *sufficient* on its own to enable word learning. Moreover, it reveals how children's ability to use different cues changes over the course of development (e.g., that children are able to use each cue in increasingly complex or difficult environments). In contrast to these experiment-based environments, however, in naturalistic environments there are usually *multiple* cues available to help children learn new words.

Moreover, individual cues are often ambiguous or insufficient – helping children eliminate some, but not all possible referents.

Subsequent theoretical work on word learning has sought to integrate the different theories described above (i.e., Constraint-Based vs. Social-Pragmatic vs. Associationist). Rather than viewing word learning through an either/or framework (children either use linguistic cues or social cues), this body or research attempts to answer questions about when and why children use different types of word-learning cues (Hirsh-Pasek & Golinkoff, 2008). A prominent integrative theory of word learning is the Emergentist Coalition Model (ECM) by George Hollich, Kathy Hirsh-Pasek, Roberta Golinkoff and colleagues (2000). The ECM adopts a hierarchical perspective towards word-learning cues: children use multiple cues, but they use different cues for word learning at different points in development. This perspective was motivated by research examining word learning in young infants (Woodward, 2004). Since very young infants are unable to use word learning cues like ME and speaker gaze, many developmental psychologists believed that the earliest stages of word learning were guided by naïve attentional mechanisms. With development, however, infants begin to use more reliable cues (like ME or speaker gaze) and realize that perceptual cues (like object salience) are not reliable (Hollich et al., 2000).

This hierarchical framework has led to a focus on how children use multiple word-learning cues; the vast majority of the work, however, has involved competing cues. By putting cues in conflict, researchers can determine which cues children will preferentially use over others (e.g., whether they will use a social cue over a linguistic cue) and whether this changes with age and language experience. Much of the research within this competitive framework has focused on two cues: speaker gaze and ME. In these experiments, children are shown a novel and familiar object. When the adult points to or looks at the familiar object while requesting an

object with a novel name, 3- and 4-year-olds are more likely to choose the novel object, rather than the familiar object, as the referent of a novel word (Jaswal & Hansen, 2006). Thus, when speaker gaze suggests one referent (i.e., the familiar object) and ME suggests a different referent (i.e., the novel object), children use ME. However, when the adult *both* points to and looks at the familiar object, 2- and 4-year-olds are more likely to choose the familiar object as the referent of the novel word (Grassman & Tomsello, 2010). On the surface, these results may appear discrepant, but likely occur because children are better able to identify the referent of a novel word using both gaze and pointing, rather than just gaze or just pointing (Booth, McGregor & Rohlfing, 2008; Jaswal & Hansen, 2006). This suggests that children's preference for one cue over another is not absolute or set in stone; whether or not children will use ME depends upon the strength of the social cues that are available as an alternative.

The relative weight that children assign to social cues vs. ME also changes based on children's language background. Children who are learning more than one language more frequently violate ME, because they are learning multiple labels (one in each language) for each object. Although 17- to 22-month-old monolinguals use ME to correctly identify the referent of a novel word, their multilingual peers do not (Houston-Price, Caloghiris & Raviglione, 2010; Byers-Heinlein & Werker, 2009). Bilingual children's willingness to use ME depends on the extent of overlap between their lexicons. 17- to 18-month-old bilinguals with significant overlap (i.e., they know the translation equivalents for over half of the words in their vocabulary) do not use ME, while bilinguals with less overlap use ME to identify the referent of a novel word like their monolingual peers (Byers-Heinlein & Werker, 2013).

With less reliance on linguistic cues like ME, bilingual children rely more on social cues. 3- and 4-year-old bilinguals are better able than their monolingual peers at identifying the target

of a speaker's gaze (Yow & Markman, 2011). When they must use a speaker's gaze to identify the referent of a novel word, 4-year-old multilinguals are more successful in learning and remembering the novel word-object mappings than their monolingual peers (Yow, Li, Lam, Gliga, Chong, Kwek & Broekman, 2017). Multilingual children's increased reliance on social cues and decreased reliance on linguistic cues, however, is not global, but rather context specific. When social cues and linguistic cues are put into conflict, 3- and 4-year-old bilinguals are more likely to use ME (i.e., choose the novel object that the speaker was *not* pointing to) if the speaker had previously read a story in only one of the child's languages compared to both of the child's languages (Hung, Patrycia & Yow, 2015). These findings suggest that the weight bilinguals assign to social vs. linguistic cues changes based on their conversational partner. Multilingual children are not necessarily more likely to relax the ME constraint in all environments, but specifically in multilingual environments.

This last possibility – that the relative weights children assign to different cues may change not only across development, but also across environments – is not restricted to multilinguals. In particular, the weights that all children assign to different cues might vary based on the reliability of each cue in that environment. Such re-weighting occurs in perceptual tasks that put auditory and visual cues or haptic and visual cues in conflict against one another. Visual dominance usually occurs, because the variance in visual estimates are lower than audio or haptic estimates. When the variance of visual estimates increases, however, (e.g., when an image becomes blurrier) then visual dominance disappears (Alais & Burr, 2004; Ernst & Banks, 2002). Adults also adjust the relative weights that they assign to different word-learning cues. When learning new words, adults track cross-situational statistics less as the reliability of a speaker's gaze improves (MacDonald, Yurovsky & Frank, 2017). To date, however, little research has

examined how the relative weights that children assign to different word-learning cues may be context-dependent.

By putting cues in conflict, a growing field of research has revealed the relative weights that children assign to word learning cues. Moreover, we have a better understanding of how these weights may vary based on children's cognitive development, their language experience (mono- vs. multilingual), and even the local linguistic context. Recent experimental work, however, has failed to find evidence supporting a weighted cue combination approach (Yurovsky & Frank, 2015). Instead, changes in children's word learning appear to be driven more by domain-general improvements in attention, memory, and processing speed (e.g., Fernald et al., 1998). Moreover, in contrast to what hierarchical frameworks like the ECM propose, older children are still influenced by unreliable "word-learning cues" like objects' visual salience. Familiar objects with high salience interfere with 3-year-olds' ability to use ME to learn new words (Pomper & Saffran, 2018).

Cues in Cooperation

Regardless of whether children preferentially weight one cue over another, children may still benefit from the presence of *both* cues. When two cues conflict in identifying the referent of a novel word, children must use one cue over another. When the same two cues *cooperate*, however, children do not need to choose one cue, but rather can use both cues. Indeed, both monolingual and bilingual 4- and 5-year-olds are more successful in learning new words when ME and speaker gaze cooperate in identifying the same referent of a novel word compared to when both cues compete (Gangopadhyay & Kaushanskaya, 2020). To date, however, very little research has systematically examined whether and how children use multiple cooperating cues in

word learning. Despite the paucity of research, many developmental psychologists acknowledge that the presence of multiple cooperating cues plays an important role in novel word learning:

> From the beginning, then, word learning is multiply constrained. Even young 1-year-olds draw on both default assumptions [like ME] and behavioral cues [like speaker gaze] in word learning…Any account proposing a single factor is responsible for early word learning will be lopsided at best. Because word learning in the wild most likely involves multiple constraints, it may not be a clear reflection of any one of them. Experiments are useful for clarifying the role of individual constraints on word learning, but they do not always shed light on the ways in which multiple constraints converge in natural contexts (Woodward, 2000; p.106).

Indeed, results from several of the experiments described above suggest that the presence of multiple, cooperating cues may not be merely beneficial, but rather *necessary* for word learning in some contexts. 12-month-olds are only successful in learning and retaining novel word-referent mappings when the speaker both looks at and holds the target object, but not when the speaker just looks at the object (Hollich et al., 2000). Similarly, 2-year-olds are only successfully in learning and retaining novel word-referent mappings when ME was combined with the speaker holding the object (Horst & Samuelson, 2008). Even when an individual cue is sufficient, the presence of multiple cues may provide an additive benefit. While 28- to 31-month-olds successfully use speaker gaze to learn new words, they are significantly more accurate in learning and retaining the word-referent mappings when gaze was supplemented with pointing or touching (Booth, McGregor & Rohlfing, 2008).

Taken together, these experiments reveal that the addition of pointing or holding improves children's success in word learning (even enabling word learning in situations where children would otherwise fail). However, it remains unclear whether the added benefit was due to the presence of multiple cues or simply the addition of a social cue that is a strong and reliable indicator of a speaker's referential intent (i.e., the speaker pointing to or holding the object). This is particularly important, in light of the hierarchical framework in theories of word learning like ECM and the debate about whether social cues are more important than linguistic cues or vice versa (e.g., Tomsello, 2000). Prior research is unable to discern between the different ways in which multiple, cooperating cues may affect children's accuracy in word learning.

One possibility is that the addition of a second cue will always improve word learning. Regardless of the type of cues available, having two cues may always be better than having just one. Computational work in related areas of language acquisition suggest that this might be the case. The presence of multiple cues improves the speed and accuracy with which computational models (simple recurrent networks) can be trained to segment words from continuous speech and track the cross-situational co-occurrence between different words and labels (Christiansen, Allen & Seidenberg, 1998; Monaghan, 2017). Similar additive benefits have been found with computational work involving other cognitive tasks, including perception and spatial memory (Ernst & Banks, 2002; Xu, Regier & Newcombe, 2017).

A second possibility, however, is that only the addition of some, but not all, cues will improve word learning. This possibility is consistent with hierarchical models of word learning like the ECM (Hollich et al., 2000). If there is indeed a hierarchy between cues, when a preferred cue is available it may be used instead of other (dispreferred) cues. Dispreferred cues will only be used if the preferred cue is absent. For instance, adults are less likely to track cross-situational

statistics (a more effortful cue that is distributed across learning experiences) when the reliability of a speaker's gaze (a less effortful and immediate cue) increases (MacDonald, Yurovsky & Frank, 2017). Similarly, infants are more likely to check a speaker's gaze only when there is referential ambiguity (i.e., there are two novel objects present; Baldwin & Tomasello, 1998). It is plausible that children's word learning will benefit when ME is added to speaker gaze, but not vice versa. As previously described, children struggle to use only a speaker's gaze to learn new words (Booth, McGregor & Rohlfing, 2008; Jaswal & Hansen, 2006). It is equally plausible, however, that the exact opposite will be true – children's word learning may only benefit when speaker gaze is added to ME, but not vice versa. Prior research has found that younger children are only able to learn new words when ME is augmented with social cues (Horst & Samuelson, 2008).

Orthogonal to these alternatives is the possibility that the presence of multiple cues enables word learning that is robust to variation or noise. In all of the experiments described above, individual cues were sufficient for word learning. For instance, in experiments testing ME there may be multiple familiar objects present, but there is only ever *one* novel object present. Similarly, in experiments testing speaker gaze, the adult unambiguously looks at *one* of the novel objects that are present. In natural word-learning environments, however, individual cues are often ambiguous or unreliable predictors of the intended referent. There are usually multiple unfamiliar objects present and a speaker's gaze may be in the general vicinity of several objects. In some instances, an individual cue may even be misleading or absent (e.g., a speaker may be looking at the wrong object or their eyes may not be visible). Rather than solely relying on one cue, the best way to determine referential intent is to aggregate over multiple, overlapping cues (Frank, Tenenbaum & Fernald, 2013).

The presence of multiple, cooperating cues may be critically important for word learning, because it creates degeneracy, rather than redundancy. Degeneracy is "the ability of elements that are structurally different to perform the same function or yield the same output" (Edelman & Gally, 2001, p. 13763). Unlike redundancy, degeneracy does *not* imply that the elements are fully equivalent. Just as different proteins perform overlapping biological functions, different word-learning cues perform overlapping functions (i.e., helping children correctly identify the referent of the novel word). This degeneracy may be critically important for word learning (Monaghan, 2017). Systems with degeneracy are robust to variation or noise in elements because there is a reduced dependency on individual elements. In the same way that biological systems can function in the absence of a protein (e.g., due to a knock-out gene), children can successfully learn words in the absence of a cue (e.g., when a speaker's gaze is unavailable). This feature – robustness to noise – may be critically important if we seek to shift our understanding of word learning from the laboratory to the real world. By carefully examining the effect of individual cues on novel word learning, past research has potentially removed one of the most important properties of these cues: their degeneracy.

Overview of Current Research

The current research systematically examines how the presence of multiple, cooperating cues affects children's novel word learning in two different contexts:

1. When individual cues are sufficient, does the presence of multiple cues improve children's accuracy in referent selection and word learning? (Study 1)
2. When individual cues are insufficient, does the presence of multiple cues enable word learning? (Study 2)

Each study consisted of a different sample of 3-year-olds. In both studies, children watched videos of an adult labeling novel objects in the presence of one or more distractor objects. In Study 1, different combinations of cues were available to help children identify the referents of the novel words: for some trials, the referent could be identified by using the speaker's gaze, for other trials the referent could be identified by using ME, and for other trials the referent could be identified using both the speaker's gaze and ME. In all three conditions, individual cues were sufficient – eliminating all distractors and unambiguously identifying which object was the referent of the novel word. In Study 2, the correct referent could only be identified through a combination of both the speaker's gaze and ME. Individual cues were insufficient – eliminating some, but not all distractors.

Immediately after the teaching videos for each study, children's learning and retention of the word-object mappings was measured. On each test trial, children saw images of two novel objects from the videos and heard a sentence identifying one by name. In the absence of any cues, the target novel object could only be identified if children had successfully learned and retained the word-object mappings from the teaching videos.

Children's eye movements were tracked during both teaching and test trials. In Study 1, children's accuracy in fixating the target novel object was compared between conditions to examine whether the addition of a second cue always, sometimes, or never improves children's word learning. These comparisons between conditions were made for both teaching and test trials. We predicted that the addition of a second cue would always improve word learning. Specifically, children would be more accurate in identifying the referent of a novel word on teaching trials with both cues (Gaze+ME) compared to teaching trials with just one cue (Gaze or

ME). Similarly, we predicted that children would be more accurate on test trials for novel word-object pairings from teaching videos that included both cues (Gaze+ME) compared to just one cue (Gaze or ME). In additional analyses, we examined whether there was a correlation between children's accuracy on teaching and test trials. We predicted that children who were more accurate in fixating the target object during teaching would also be more accurate at test. In exploratory analyses, we examined whether condition differences in children's accuracy on teaching trials resulted from increases in children's fixations to the distractor object or to the speaker's face. In Study 2, where there was only one condition, children's accuracy in fixating the target novel object during test trials was compared against chance. We predicted that children would succeed when word learning required the combination of two cues. Specifically, children's accuracy in fixating the target novel object on test trials would be significantly higher than chance, which is 50% (i.e., equal fixations to both the target and distractor).

Study 1

Method

Participants. The final sample included 51 children (30 female, 21 male) with an average age of 39.7 months (range: 38-42 months). This age range was selected based on prior research, because it is the earliest age at which children are able to learn and retain novel words using only a speaker's gaze or only ME (Yurovsky, Wade & Frank, 2013; Bion, Borovsky & Frank, 2013). Children were recruited from a database of interested families living in or near a mid-sized city in the Midwestern United States. The demographics of the final sample included 45 children who were Caucasian, five who were Caucasian and Asian, and one who was Caucasian and Pacific Islander. One child was Hispanic/Latino. All children were reported to be monolingual English-

learners (hearing no more than 10 hours per week of languages other than English), to have normal or corrected-to-normal and hearing vision, be born no more than 4 weeks before their due date, and without parental concerns regarding their speech/language development. An additional thirteen children were tested but not included in the final sample, because they ended the experiment early (n=4), had too much missing data (n=8; see criteria below), or because of parental concerns about hearing (n=1). All parents provided written informed consent and children provided oral assent. The experimental protocols, including procedures for obtaining informed consent, were approved by the local IRB.

Procedure. Participation involved a single visit lasting approximately 45 minutes. Children and their parents were first brought into a waiting room with toys. Here, children were allowed to acclimate to the lab environment, while their parents completed the consent process and the experimenter answered any questions. After children provided verbal assent, both the child and caregiver were led into a sound-proof booth where the child completed the word learning task. Afterwards, caregivers filled out demographic and background questionnaires in the waiting room. The experimenter provided a debriefing and answered any additional questions. Children then picked out a thank you gift for participating.

Word learning task. In a sound-proof booth, children were seated on their caregiver's lap approximately 2 feet away from a 55-inch TV and 60 centimeters away from a Tobii X2-60 eye tracker that was mounted underneath the TV. A low-light video camera was also mounted below the TV. All caregivers wore opaque sunglasses to prevent them from seeing the visual stimuli and to prevent the Tobii from tracking their eye movements. Caregivers were instructed to help keep children seated on their lap and centered in front of the eye tracker. Before the start of the word learning task, children completed a 5-point calibration procedure. If calibration was poor

(i.e., the eye tracker was unable to reliably track children's fixations for at least 3 points), the procedure was repeated. Poor calibration can result from a child's inattentiveness or the eye tracker being unable to reliably detect the child's pupils (e.g., because the child moved out of range). If calibration was still poor following the second repetition, the experiment was still administered.

Stimuli. Images of the six novel objects and novel labels were selected from the Novel Object & Unusual Name Database, 2e (NOUN; Horst & Hout, 2016). Six novel objects were chosen that were rated by adults to be minimally familiar and maximally different in visual similarity (Horst & Hout, 2016). Three monosyllabic and three disyllabic novel words were chosen to be maximally different phonologically – no two words shared the same consonant onset or consonant/vowel offset. The same object-label pairings were used for all children (see Figure 1). Objects and labels were yoked into three pairs; in each pair, one label was monosyllabic and the other label was disyllabic. The assignment of novel word-object pairs to each condition was counterbalanced between children.[1] Twenty familiar objects were selected from the MacArthur-Bates Communicative Development Inventory (CDI). These objects were chosen such that their labels were highly familiar to 30-month-old children (the oldest possible age for the CDI) based on CDI norms from the Wordbank database (Frank, Braginsky, Yurovsky, & Marchman, 2017). Images of the familiar objects were found online. Additional images of three novel objects were selected from the NOUN database to be used as distractors on trials where familiar objects were labeled (see below). All images were edited using Photoshop

[1] The effect of condition, therefore, cannot be idiosyncratic to any word-object pairing (e.g., the results are not affected if a specific novel word-object pairing was easier to learn, because that pairing occurred equally often in each condition across children).

so that objects were approximately matched in size and visual salience and placed on an empty background.

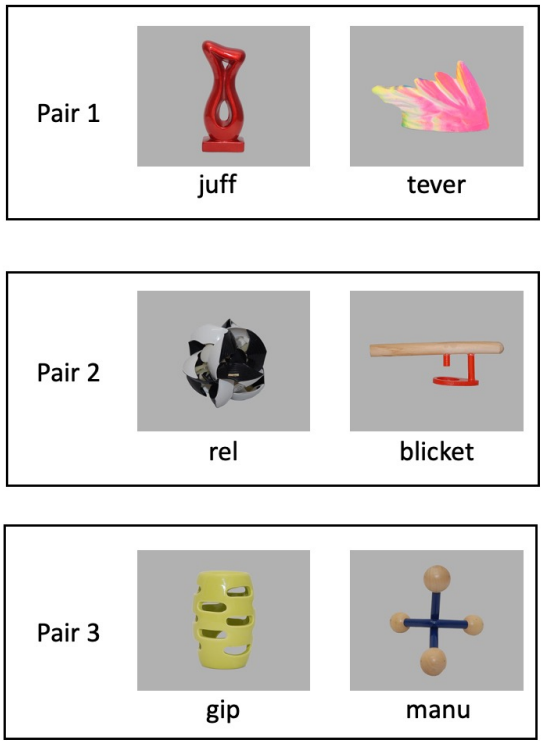

Figure 1. Novel objects and their associated labels used in Study 1.

Teaching trials. Each trial consisted of a video of an adult seated at a table with two objects. One object was positioned on the table to the left of the adult, the other object was positioned on the table to the right of the adult.[2] At the beginning of each video, the adult looked into the camera, then looked at the object on the right side of the table (from the viewer's

[2] Objects were not physically present on the table, but rather digital images of the objects that were super-imposed using video-editing software (iMovie). This enabled more effective counterbalancing (i.e., the same video of the adult was used for trials with different objects and different conditions) and reduced the amount of generalization necessary, because the same images of the novel objects were used on test trials.

perspective), the object on the left side of the table, and then returned her gaze to the camera. Finally, she labeled one of the objects (e.g., “It’s a flower”).

Which objects were present on the table and where the speaker looked immediately after she labeled the target object varied between trials (see Figure 2 for still frames of videos from each Condition). On *Familiar* trials, the target was a familiar object with a known name and the distractor was a novel object. The speaker looked at the target familiar object immediately after labeling it. On *Gaze* trials, both the target and distractor were novel objects. The speaker, however, looked at the target novel object immediately after labeling it. On *ME* trials, the target was a novel object and the distractor was a familiar object with a known name. The speaker, however, did *not* look at the target novel object after labeling it, but instead looked at the middle of the table between both objects. On *Gaze+ME* trials, the target was a novel object and the distractor was a familiar object with a known name. The speaker looked at the target novel object immediately after labeling it.

All videos were recorded in high definition (1920 x 1080 pixels) using a camera mounted on a tripod. Multiple recordings were made of each trial. A recording was selected for each trial that maximized similarity in timing across trials. Selected recordings were then edited using iMovie to trim the beginnings and ends of each video so that all recordings had the same total duration of 7,700 ms. On average, the speaker spent 1,400 ms looking into the camera at the beginning of each video, then 3,300 ms looking at the objects, 1,400 ms labeling the target object, and then 1,600 ms fixating either the target object or the middle of the table.

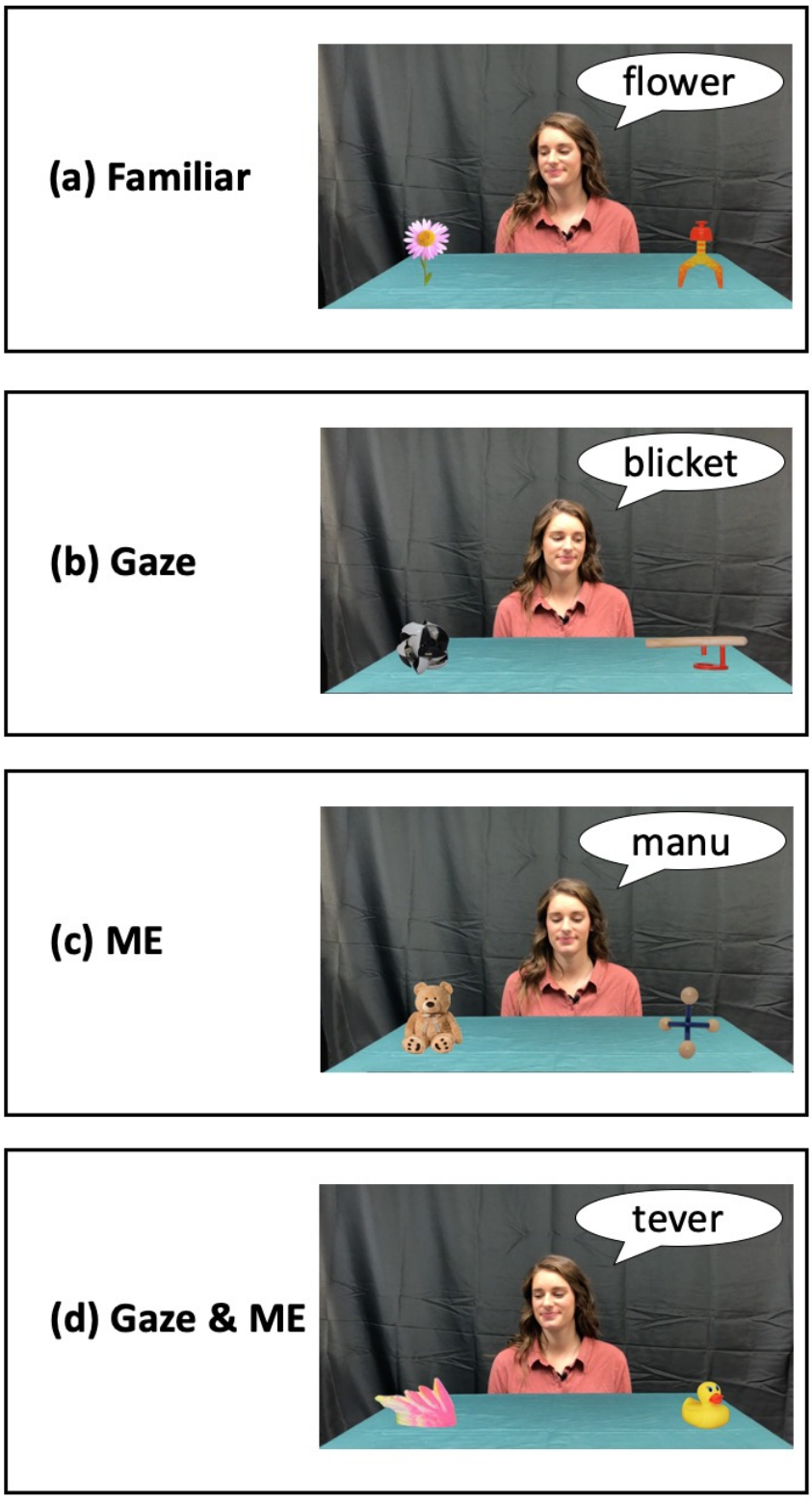

Figure 2. Still frame from a teaching videos in each Condition for Study 1. Frames are taken immediately after the offset of the target word (displayed in the text bubble).

Test trials. On each trial, children saw images of two objects – one image was displayed in the bottom left corner of the screen, the other image in the bottom right corner. The images were of the same familiar and novel objects from the teaching trials placed on gray backgrounds. For trials with novel objects, both objects were from the same condition (e.g., both were labeled on Gaze+ME teaching videos).[3] Children then heard a sentence labelling one of the objects by name followed by a generic sentence to maintain their attention (e.g., "Where's the modi? That's cool!"). These sentences were produced by the same adult in the teaching videos. Multiple recordings were made for each trial. A recording was selected for each trial that maximized similarity in timing and intonation contour across trials. Recordings were then edited using Praat to match the duration of each segment across trials. Additionally, the amplitudes of each recording were scaled to 65 dB. Each trial lasted 5,505 ms, beginning with 1,500 ms of silence, a 758 ms carrier phrase, a 576 ms target word, another 1,000 ms of silence, a 1,071 ms generic phrase to maintain attention (e.g., "That's cool!"), and ending with 600 ms of silence.

Trial order. The entire experiment consisted of 36 trials in total. This included 15 teaching trials and 21 test trials. The six novel objects were each the target on two teaching trials. Three familiar objects (juice, book, flower) were each the target on one teaching trial. For teaching trials in the Gaze condition, the distractor was the other novel object in the yoked pair. For teaching trials in the ME and Gaze+ME, the distractors were familiar objects with known names (each familiar object was only used on one trial). The six novel objects were each the

[3] This was necessary to prevent potential learning during test trials. If children learned the names of novel objects in one condition, these objects would become, in a sense, familiar objects with known names. If these objects were used as distractors, children could then use ME to learn the names of the novel objects from other conditions during testing. Additionally, if a novel object was paired with different novel objects on multiple trials, children could track the cross-situational statistics to learn the names of the novel objects.

target on three test trials and three familiar objects (dog, boat, cat) were each the target on one test trial. Different familiar objects were used on teaching and test trials in order to minimize repetition and maintain children's interest. For test trials with novel objects, the distractor was the other novel object in the yoked pair. For test trials with familiar objects, the distractors were other familiar objects (bird, towel, horse). Since two novel objects were assigned to each condition (Gaze, ME, Gaze+ME), there were four teaching trials per condition and six test trials per condition.

Trials were arranged into three blocks by condition. Therefore, children saw every teaching trial and then every test trial for a given condition, before they saw teaching and test trials for the other conditions (see Appendix for an example of a trial order). The first teaching trial and the first test trial in each block always had a familiar object as the target. This was done in order to acclimate children to the procedure at the onset of the experiment and to ensure that familiar objects were *not* always distractors (a regularity that children could readily detect and use to fixate the target object before it was labeled). Finally, the assignment of each condition to each block was counterbalanced between children. Therefore, some children started with the Gaze condition, others with the ME condition, and yet others with the Gaze+ME condition.

Trials were arranged in a pseudo-random order. The target was never on the same side of the screen for more than three consecutive trials. The target was never the same object for more than two consecutive trials. The location of targets and distractors was balanced as closely as possible given the odd number of trials. On teaching trials, the target occurred eight times on the left and seven times on the right. On test trials, the target occurred 11 times on the left and 10 times on the right. These constraints were necessary to make sure that children did not notice any regularities in trial structure that they could use to fixate the target before it was labeled.

Data collection. Children's eye movements were tracked using a Tobii X2-60 eye tracker. The Tobii X2-60 tracks the location (in x- and y-coordinates) of children's fixations every 16 ms. Regions of Interest (ROIs) were determined separately for teaching and test trials. Teaching trial ROIs were: the target object, distractor object, and the speaker's face. Screen shots were taken from each teaching trial, rectangles were drawn to cover each ROI using Photoshop, the coordinates were then averaged across all rectangles to determine the average x- and y-coordinates for each ROI. Test trial ROIs were the target and distractor images; the exact x- and y-coordinates for these ROIs were known, since the image locations were determined via python.

Unless otherwise specified, the dependent variable for all analyses is the proportion of time children spent fixating a specific ROI (e.g., the target object) out of the total time spent fixating all ROIs (e.g., the target object, distractor object, and face). Frames for which children were not fixating any ROI or children's fixation location could not be tracked (due to blinking, pointing, tilting of the head, etc.) were coded as NAs and treated as missing data. Proportions were calculated during a critical window 300 to 1,800 ms after the onset of the target word on each trial. This is the standard window used for most looking-while-listening paradigms, because it takes children approximately 300 ms to program an eye movement and around 1,800 ms their attention begins to wane (Fernald et al., 2008). Therefore, fixations before and after this window were not considered to be stimulus-driven (i.e., in response to the target word). For teaching trials, this 1,500 ms critical window occurred on average between 5,833 to 7,333 ms from the onset of the trial. The onset of the target word, however, varied from trial to trial. Therefore, the critical window was adjusted individually for each trial. For test trials, the 1,500 ms critical window occurred between 2,558 to 4,058 ms from the onset of the trial. This window was the same for all test trials, because the auditory stimuli were edited to exactly match in duration.

Data cleaning. Individual trials were excluded from analyses if the child was not fixating any of the ROIs for more than half (750 ms) of the critical window. These trials were excluded, because there was too much missing data. After cleaning, we identified 28 children who had fewer than two useable teaching or test trials in one or more conditions (e.g., a child with only 1 useable teaching trial for the Gaze condition would be flagged). Seven of these children were excluded from the final sample, because they were identified as inattentive throughout the procedure by the experimenter. Video recordings for the remaining 21 children were hand coded. Using custom software, coders determined for each frame whether the child was looking at the left image, right image, speaker's face, or away. Coders were unaware of the target object, target location, and condition for each trial. After hand coding, only one child still had too much missing data and was therefore excluded. Tobii and handcoded data were combined by downsampling the Tobii data from 60 Hz to 30 Hz by binning every 33 ms and calculating children's average fixation proportion within each bin.

Children in the final sample had on average 3.73 (SD=0.5) useable teaching trials (out of the maximum of 4) and 4.52 (SD=1.3) useable test trials (out of the maximum of 6). There was no difference in the number of useable teaching trials between conditions, $\chi^2(2) = 4.2$, $p = 0.12$. There was, however, a marginally significant difference in the number of useable test trials between conditions, $\chi^2(2) = 5.7$, $p = 0.06$. On average, there were 4.8 useable test trials for the ME condition, 4.4 for the Gaze condition, and 4.3 for the Gaze+ME condition. Since test trials were identical across conditions (the only difference between conditions occurred on teaching trials) and the assignment of objects to each condition was counterbalanced between children, it is unclear what may have caused these differences. These differences are unlikely to affect the results, because they do not substantially reduce statistical power in one condition compared to

the others. More concerning, however, are any differences in the number of useable trials between blocks. A decrease in the number of useable trials with each successive block indicates that more trials are being excluded in later blocks and suggests that children are becoming inattentive towards the end of the experiment. There was no difference in the number of useable teaching trials between blocks, $\chi^2(1) = 2.2$, $p = 0.13$. The number of useable test trials, however, significantly decreased with each subsequent block, $\chi^2(1) = 33.8$, $p < .001$. On average, there were 5.2 useable trials in the first block, 4.4 useable trials in the second block, and 3.9 useable trials in the third block. For this reason, the main analyses involving test trials were repeated with the data restricted to the first block. This analysis may provide converging evidence, but is on its own likely underpowered (since the number of observations are significantly reduced and the manipulation of word learning cues is no longer within-subjects, but rather between-subjects).

Statistical analyses. Experimental stimuli and children's gaze locations were presented and tracked using custom python code (and the pygaze package). All analyses were carried out in RStudio (version 1.2.5001) using the lme4 package (version 1.1-21). The proportion of children's fixations to different ROIs was regressed on Condition (Gaze, ME, Gaze+ME) both for teaching and test trials. The full random effects structures were included in all models (Barr, Levy, Scheepers, & Tily's, 2013). Unless otherwise specified, all analyses were carried out using linear mixed effects models at the trial level and were fit using maximum likelihood estimation. Estimates of degrees of freedom and significance tests were completed using the Kenward-Roger procedure. When the omnibus (2 degree of freedom) test revealed a significant effect of condition, pairwise comparisons were conducted between all three conditions (i.e., Gaze vs. ME, Gaze vs. Gaze+ME, ME vs. Gaze+ME). When the omnibus test is significant and only 3

pairwise comparisons are necessary, the pairwise comparisons are Fischer LSD protected and therefore p-values do not need to be adjusted for multiple comparisons.

Results

Teaching trials. Changes in the proportion of children's fixations to each ROI (target object, distractor object, speaker's face) over the course of teaching trials are plotted in Figure 3. Before the target object was labeled, children were predominately fixating the speaker's face and were equally likely to fixate either the target or distractor object. When the speaker began talking, children shifted their attention, increasing fixations to both objects and decreasing fixations to the speaker's face. After the onset of the target word, fixations to the target object continued to increase, while fixations to the distractor object and the speaker's face decreased. All analyses of differences between conditions (Gaze vs. ME vs. Gaze+ME) focus on this final segment, during a critical window 300 to 1,800 ms after the speaker began labeling the target object (approximately 5,833 to 7,333 ms from trial onset).

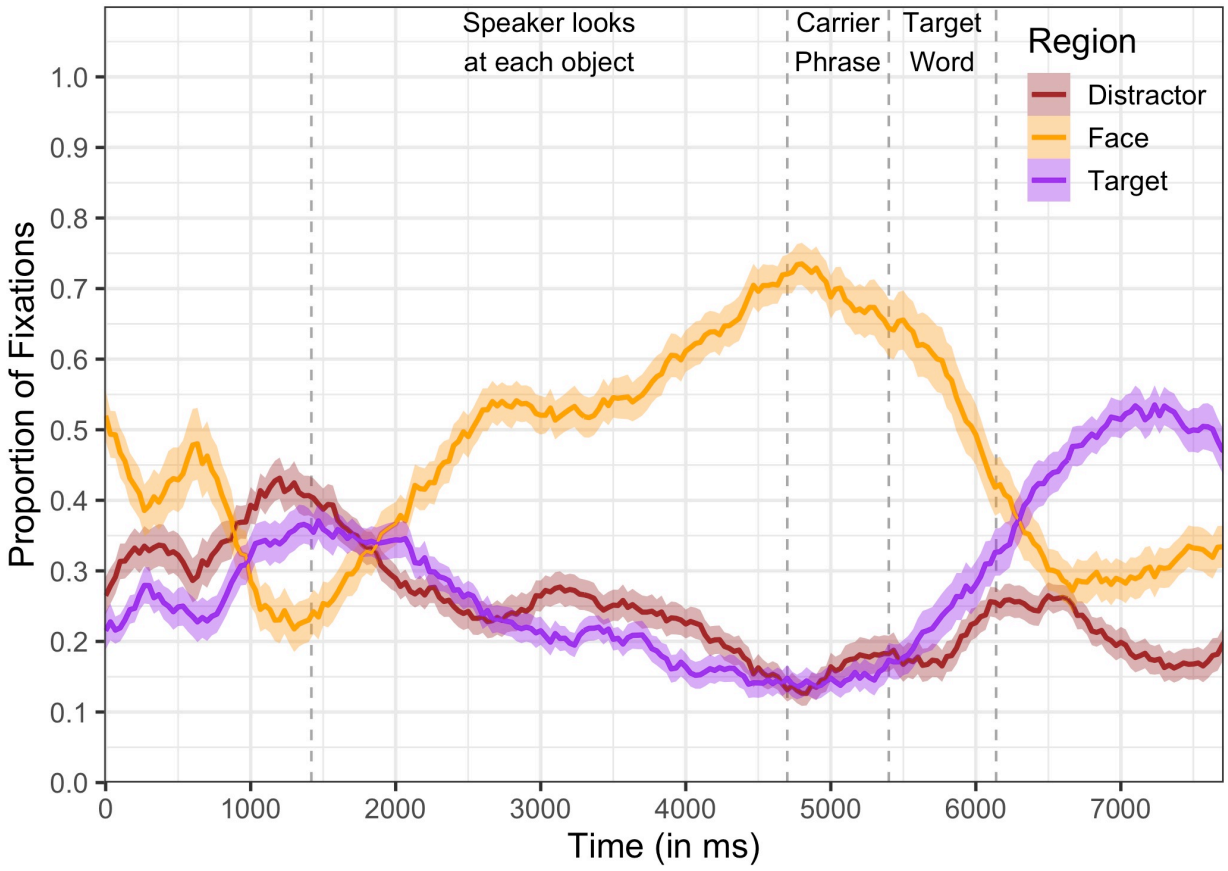

Figure 3. Time course of changes in the proportion of children's fixations to each ROI throughout the duration of teaching trials collapsing across all conditions (Gaze, ME, and Gaze+ME). Solid lines represent the proportion of fixations averaged across trials and participants. Ribbons around the lines represent +/- 1 SE. Vertical dashed lines in gray mark the average beginning and end of different behaviors during the video.

Target fixations. Our first prediction was that children would be significantly more accurate in identifying the referent of novel words when they are able to use two cues (Gaze+ME trials) compared to just one cue (Gaze trials and ME trials). Children's accuracy in fixating the target object was indeed significantly different between conditions, $F(2,48.4) = 10.3$, $p < .001$ (see Figure 4). Children's accuracy on Gaze+ME trials ($M = 49\%$, $SD = 18\%$) was significantly

higher than their accuracy on Gaze trials (M = 34.7%, SD = 20.2%), b = 0.14, $F(1,49.4) = 16.5$, $p < .001$. Children's accuracy on Gaze+ME trials was marginally higher than their accuracy on ME trials (M = 43%, SD = 18.9%), b = 0.07, $F(1,49.7) = 3.6$, $p = .06$. Finally, children's accuracy on ME trials was significantly higher than their accuracy on Gaze trials, b = 0.08, $F(1,49.4) = 5.1$, $p < .05$. Taken together, these results are only partially consistent with our first prediction. Children's accuracy in fixating the target object did increase with the addition of a second cue, but this benefit is asymmetric. The boost in accuracy was larger when ME was added to Gaze (i.e., the increase in accuracy from Gaze to Gaze+ME) compared to when Gaze was added to ME (i.e., the increase in accuracy from ME to Gaze+ME). This asymmetry occurs because children's accuracy was lower on trials with just Gaze compared to trials with just ME.

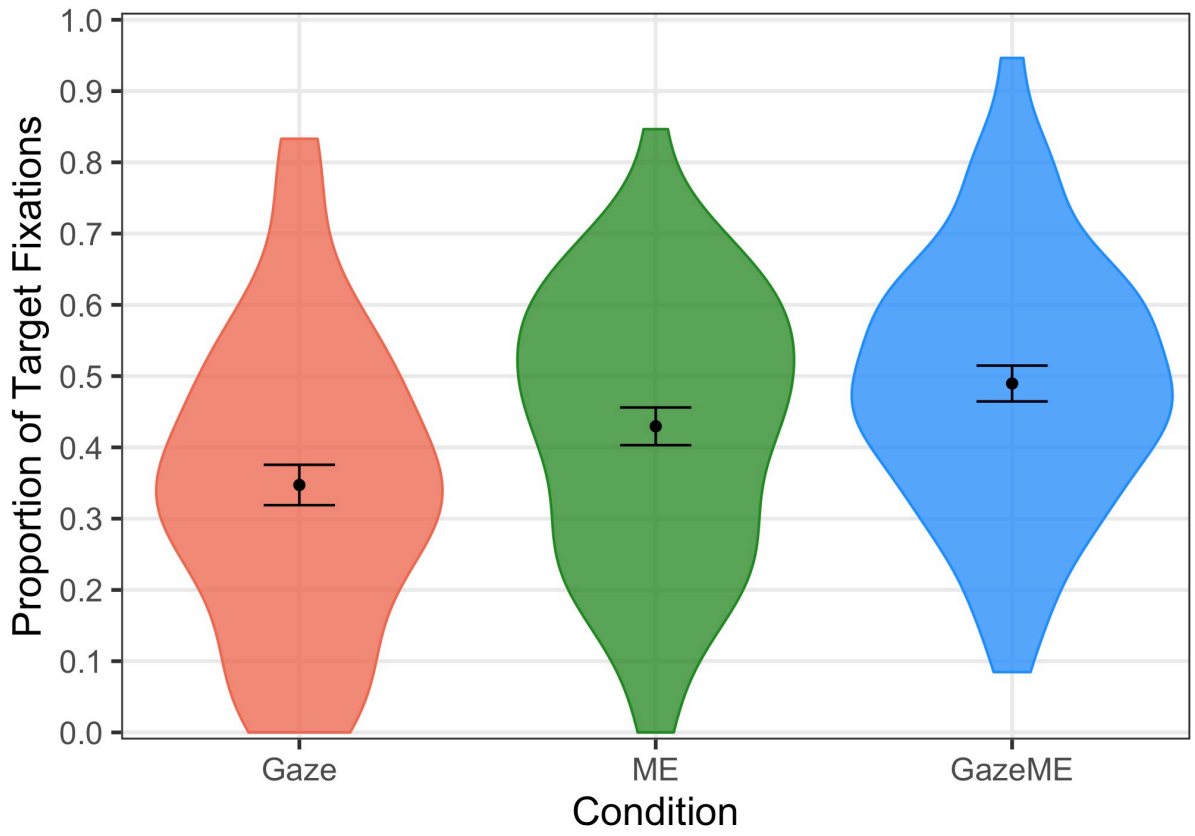

Figure 4. Proportion of children's fixations to the target object out of their fixations to the target object, distractor object, and face during a critical window 300 to 1,800 ms after the onset of the target word on teaching trials. Proportions are plotted separately by condition. Data points represent the average fixation proportion of all children. Error bars represent +/- 1 SE. Violins represent the distribution of fixation proportions across children (i.e., wider envelopes indicate more children with an average fixation proportion at that value than values with narrower envelopes).

Since there are three ROIs on teaching trials, children's accuracy in fixating the target object may differ between conditions for multiple reasons. Children may be less accurate in fixating the target object in a given condition because they are spending more time fixating the distractor, more time fixating the speaker's face, or some combination of both. In exploratory analyses, we compared proportions of children's fixations between conditions separately for the speaker's face and the distractor object.

Face fixations. As a reminder, the speaker's face was informative on Gaze and Gaze+ME trials, but not ME trials. The proportion of children's fixations to the speaker's face significantly differed between conditions, $F(2,48.5) = 17.6$, $p < .001$ (see Figure 5). The proportion of time children spent fixating the speaker's face was significantly higher on Gaze trials ($M = 46.4\%$, $SD = 25.3\%$) than ME trials ($M = 30.5\%$, $SD = 22.2\%$), $b = 0.15$, $F(1,49.3) = 28.3$, $p < .001$, and Gaze+ME trials ($M = 29.4\%$, $SD = 20\%$), $b = 0.17$, $F(1,49.5) = 26.8$, $p < .001$. Children spent an equal amount of time fixating the speaker's face on ME and Gaze+ME trials, $b = 0.02$, $F(1,49.6) = 0.29$, $p = 0.59$. This is surprising, because the speaker's face is not informative on ME trials,

but is informative on Gaze+ME trials. These results suggest that when both cues are available, children may be attending more to ME than to the speaker's gaze.

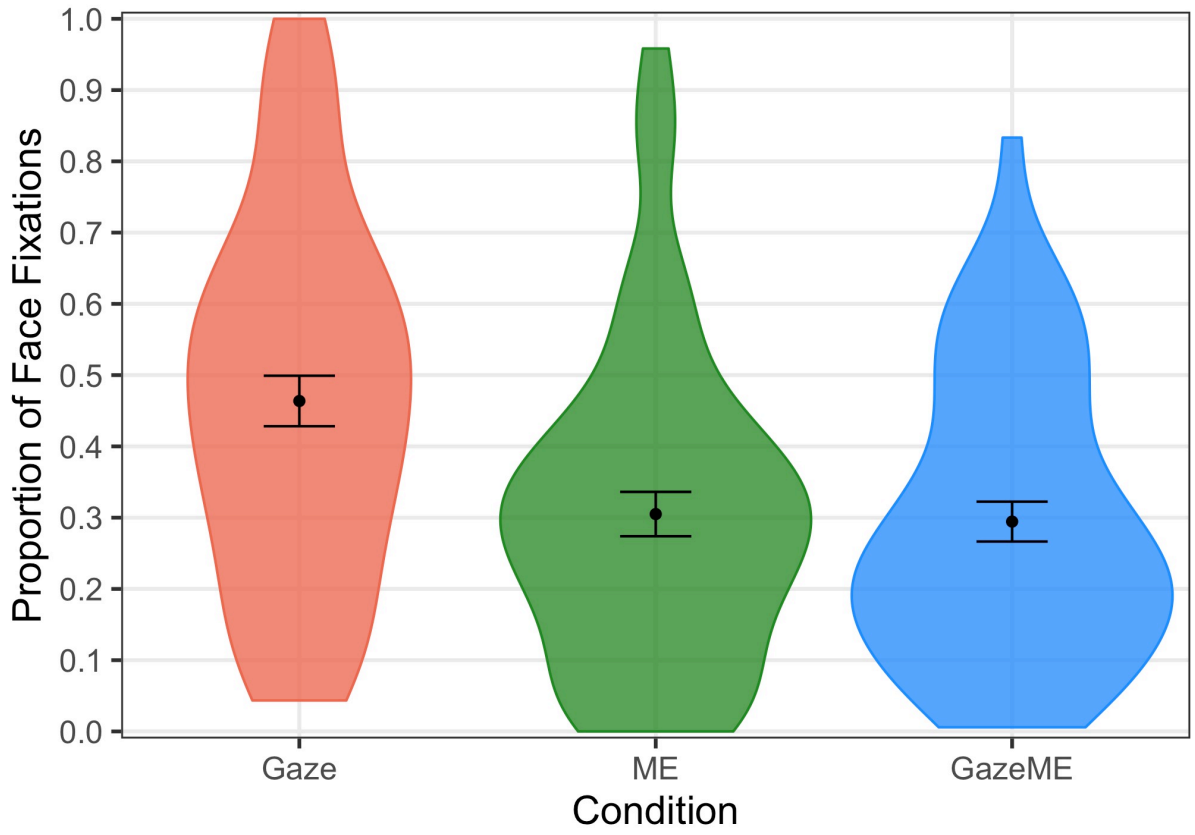

Figure 5. Proportion of children's fixations to the speaker's face out of their fixations to the target object, distractor object, and face during a critical window 300 to 1,800 ms after the onset of the target word on teaching trials. Proportions are plotted separately by condition. Data points represent the average fixation proportion of all children. Error bars represent +/- 1 SE. Violins represent the distribution of fixation proportions across children (i.e., wider envelopes indicate more children with an average fixation proportion at that value than values with narrower envelopes).

Distractor fixations. As a reminder, the distractor object was familiar on ME and Gaze+ME trials, but novel on Gaze trials. The proportion of children's fixations to the distractor object significantly differed between conditions, $F(2,48.3) = 3.7$, $p < .05$ (see Figure 6). The proportion of time children spent fixating the distractor object was significantly higher on ME trials ($M = 26.5\%$, $SD = 13.8\%$) than Gaze trials ($M = 18.9\%$, $SD = 15.7\%$), $b = 0.08$, $F(1,49.2) = 7.5$, $p < .01$. Children spent an equal amount of time fixating the distractor object on ME and Gaze+ME trials ($M = 21.6\%$, $SD = 12.9\%$), $b = 0.05$, $F(1,49.6) = 2.6$, $p = 0.11$ and on Gaze and Gaze+ME trials, $b = 0.03$, $F(1,49.2) = 1.04$, $p = 0.31$.

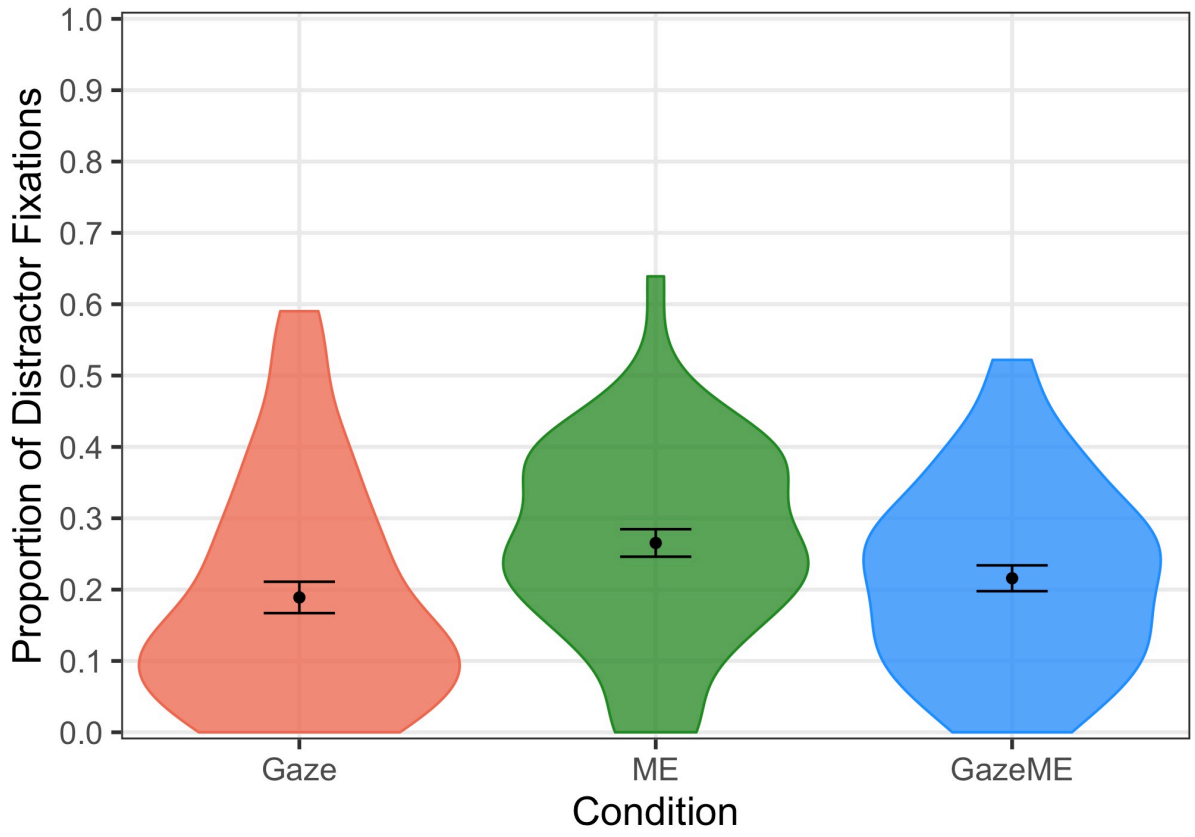

Figure 6. Proportion of children's fixations to the distractor object out of their fixations to the target object, distractor object, and face during a critical window 300 to 1,800 ms after the onset of the target word on teaching trials. Proportions are plotted separately by condition. Data points

represent the average fixation proportion of all children. Error bars represent +/- 1 SE. Violins represent the distribution of fixation proportions across children (i.e., wider envelopes indicate more children with an average fixation proportion at that value than values with narrower envelopes).

Interim summary. After hearing a novel word, children were less accurate in fixating the target object on trials where they could use only the speaker's gaze or only ME compared to when they could use both the speaker's gaze and ME. This decrease in accuracy was more pronounced when children could only rely on gaze than when they could only rely on ME. We return to this potential asymmetry in the General Discussion. These decreases in accuracy, however, occurred for different reasons. Children's accuracy in fixating the target object was lower on Gaze trials because they spent more time fixating the speaker's face than in any other condition. At the same time, however, children spent the least amount of time fixating the distractor object on Gaze trials. This is because for Gaze trials the distractor object was novel and not helpful for determining the referent, while for ME and Gaze+ME trials the distractor object was familiar and could aid children in identifying the referent of the novel word. The differences in children's accuracy in fixating the target object on trials with only ME compared to trials with both gaze and ME, however, could not be clearly attributed to any differences in fixations to the distractor object or speaker's face.

Test trials. Changes in the proportion of children's fixations to the target object out of their total fixations to the target and distractor objects over the course of test trials are plotted in Figure 7. With the exception of the Gaze+ME trials, children were equally like to fixate the

target and distractor object (i.e., 50% accuracy) at the onset of the target word.[4] For all trials, the proportion of children's fixations to the target object increased over time. Analyses of differences between conditions (Gaze vs. ME vs. Gaze+ME) focus on the critical window 300 to 1,800 ms after the onset of the target word. Interpretation of these comparisons, however, is complicated by the below chance fixations in the Gaze+ME condition, which drag down the average fixation proportion. The main analyses are therefore repeated using the average fixation proportion during the second half of the critical window (i.e., 1,050 to 1,800 ms). Finally, given the significant reduction in useable test trials towards the end of the experiment (discussed in the Method section) and the potential that earlier conditions could affect children's expectations in subsequent conditions (e.g., that the speaker's gaze is or is *not* reliable), the main analyses are also repeated using only data from the first block of trials. For these analyses, condition is a between-subjects effect (i.e., each participant only has data from one condition), rather than a within-subjects effect.

[4] Given the counterbalancing in trial orders, each trial occurred nearly equally often in each condition across children (due to participant exclusions counterbalancing was not perfect and therefore ME trials occurred on average earlier in the experiment than Gaze and Gaze+ME trials). Moreover, given the similarity across conditions (each test trial included pictures of two novel objects and a sentence labelling one object), the initial baseline differences for the Gaze+ME trials are unlikely to be the result of an idiosyncratic flaw in experimental design. Rather, these differences are likely the result of random variation or noise.

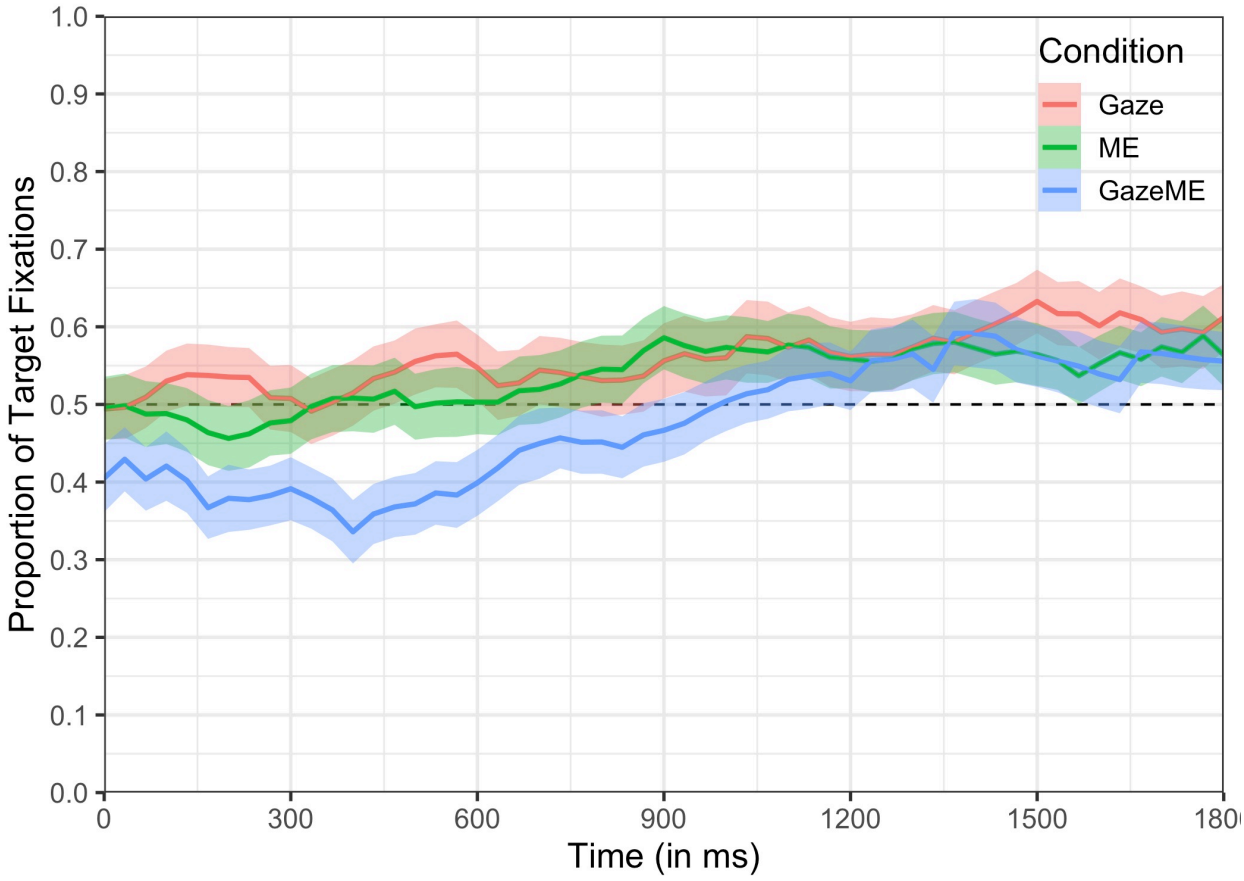

Figure 7. Time course of changes in the proportion of children's fixations to the target object out of total fixations to the target and distractor objects on test trials. Changes in children's fixation proportions are plotted separately by condition as a function of time (in ms) since the onset of the target word. Solid lines represent the proportion of fixations averaged across trials and participants. Ribbons around the lines represent +/- 1 SE. The horizontal dashed line in black represents chance (i.e., equal fixations to both the target and distractor object).

Target fixations. Our second prediction was that children would be significantly more accurate in learning and retaining the novel word-object pairings when they were able to use two cues on teaching trials (Gaze+ME) compared to when they were only able to use one cue (either Gaze or ME). Children's accuracy in fixating the target object, however, did *not* significantly

differ between conditions, $F(2,47.8) = 1.3$, $p = .27$ (see Figure 4). Children's accuracy did not differ on Gaze trials ($M = 55.7\%$, $SD = 22.9\%$), ME trials ($M = 55.4\%$, $SD = 20.9\%$), or Gaze+ME trials ($M = 49.7\%$, $SD = 18.9\%$). Collapsing across conditions, children's accuracy in fixating the target object ($M = 53.8\%$, $SD = 13\%$) was significantly higher than chance, $b = 0.06$, $F(1,48.6) = 4.2$, $p < .05$.

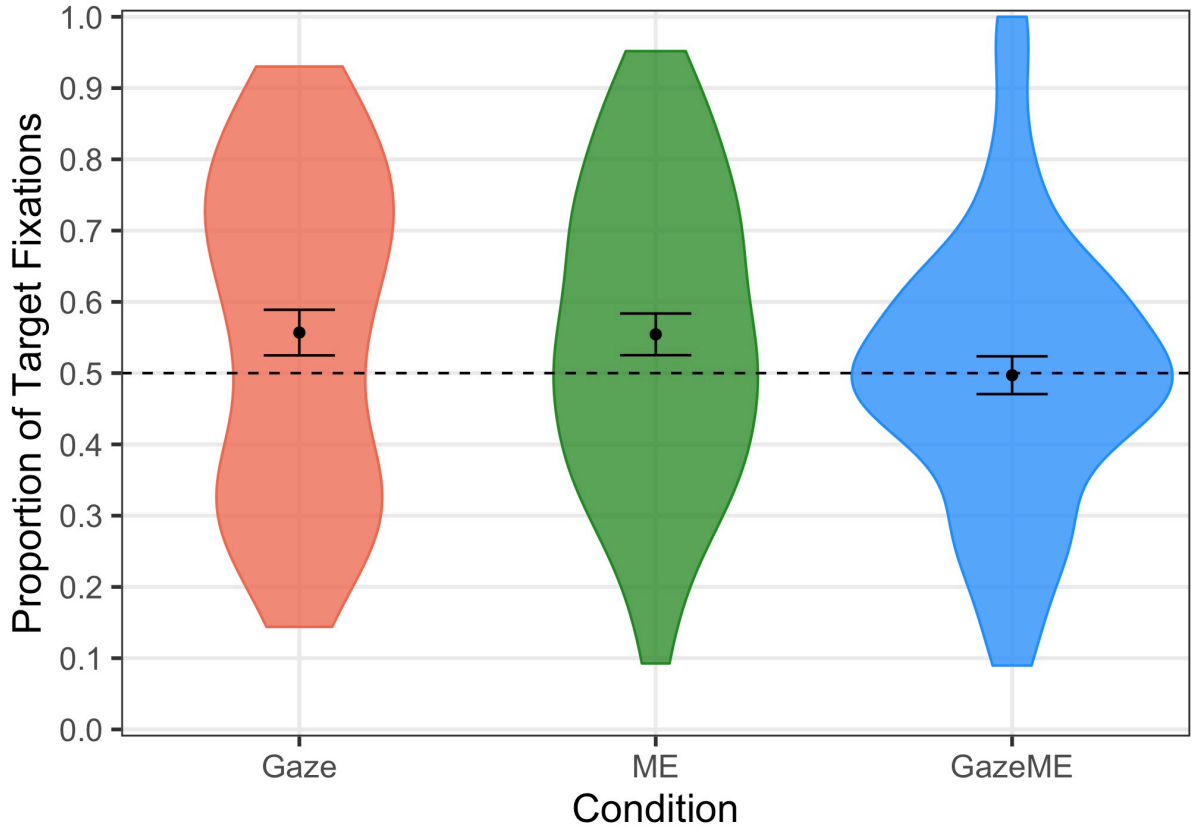

Figure 8. Proportion of children's fixations to the target object out of their total fixations to the target and distractor objects during a critical window 300 to 1,800 ms after the onset of the target word on teaching trials. Proportions are plotted separately by condition. Data points represent the average fixation proportion of all children. Error bars represent +/- 1 SE. Violins represent the distribution of fixation proportions across children (i.e., wider envelopes indicate more children with an average fixation proportion at that value than values with narrower envelopes).

Restricted window. Children's accuracy in fixating the target object during the second half of the critical window did *not* significantly differ between conditions, $F(2,47.3) = 0.2$, $p = .85$. Using this later window, the average proportions of children's fixations to the target object were higher in all conditions, particularly the Gaze+ME condition. Children's accuracy on Gaze+ME trials ($M = 54.8\%$, $SD = 22.2\%$), however, still did not differ from their accuracy on Gaze trials ($M = 57.4\%$, $SD = 25\%$) or ME trials ($M = 58.4\%$, $SD = 22.3\%$). Collapsing across conditions, children's accuracy in fixating the target object ($M = 56.6\%$, $SD = 13\%$) was significantly higher than chance, $b = 0.08$, $F(1,47.7) = 5.6$, $p < .05$.

Restricted block. When using only data from the first block, the effect of condition is between-subjects with 17 children in the Gaze condition, 19 children in the ME condition, and 15 children in the Gaze+ME condition. Children's accuracy in fixating the target object did *not* significantly differ between conditions, $F(2,47.7) = 0.8$, $p = .47$. Using data only at the beginning of the experiment, the average proportions of children's fixations to the target object were higher in all conditions. Children's accuracy on Gaze+ME trials ($M = 56\%$, $SD = 19.4\%$), however, still not differ from their accuracy on Gaze trials ($M = 64.7\%$, $SD = 23.7\%$) or ME trials ($M = 57\%$, $SD = 22\%$). Collapsing across conditions, children's accuracy in fixating the target object ($M = 61.2\%$, $SD = 23.4\%$) was significantly higher than chance, $b = 0.15$, $F(1,53.1) = 7.2$, $p < .01$.

Interim summary. Contrary to our hypothesis, children were equally successful in learning and retaining the novel word-object pairings in all conditions. These findings were robust (persisting across multiple means of analysis) and particularly striking, because children's accuracy did differ between conditions during teaching trials. Despite being more accurate in fixating the target object on teaching trials with two cues compared to trials with just one cue,

children were no more accurate in learning and retaining the name of the target object. It could be argued that children were in fact *less* accurate in learning and retaining the name of the target novel object when they could use two cues (although this is likely an artifact due to baseline differences in accuracy at the beginning of the critical window). These findings suggest that the current manipulations of children's attention during teaching trials did not affect learning outcomes. A more direct test of this hypothesis is provided in the next section, which examines whether individual differences between children in their accuracy on teaching trials were correlated with individual differences in their accuracy on test trials.

Individual differences. Our final prediction was that there would be a correlation between children's accuracy on teaching and test trials. Specifically, children who were more accurate in fixating the target object on teaching trials would also more accurate in fixating the target object on test trials. Both when collapsing across conditions and testing each condition separately, children's accuracy in fixating the target object on test trials was *not* significantly correlated with their accuracy in fixating the target object on teaching trials, p's > 0.24. Visual inspection of the correlations (see Figure 9) strongly suggests that there is indeed no relation and that these null results are not due to a lack of statistical power.

Interim Summary. Taken together, the results from Study 1 demonstrate that the addition of a second cue may improve children's ability to identify the referent of novel words but does not improve word learning. We cannot conclude, however, that the presence of multiple cues is *never* beneficial for word learning. In Study 1, children only needed one cue to identify the referent of a novel word, so the presence of a second cue was redundant. When individual cues on their own are insufficient, the presence of a second cue is *not* redundant and could enable word learning that would otherwise be impossible. We turn to this situation in Study 2.

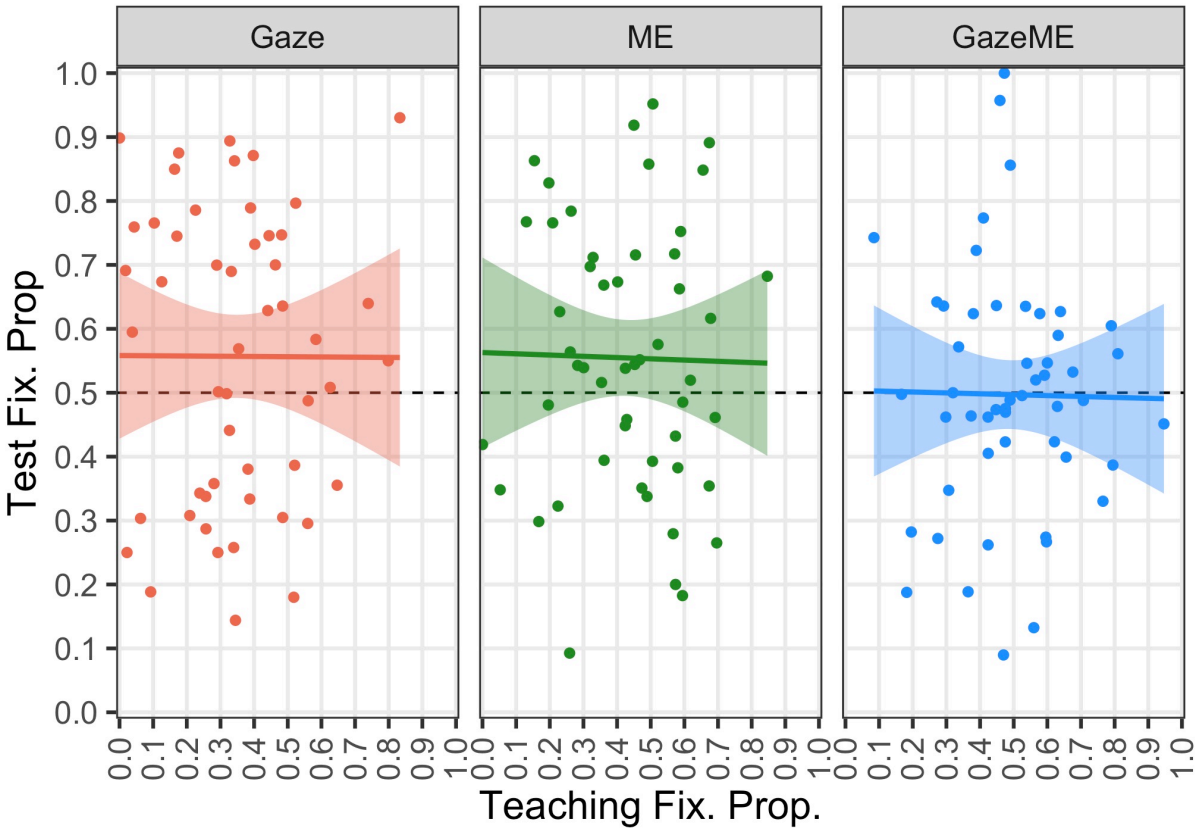

Figure 9. Proportion of children's fixations to the target object on test trials plotted as a function of their proportion of fixations to the target object on teaching trials. Data points represent the average fixation proportion for each participant. Solid lines represent the linear best fit. Ribbons around the lines are +/- 1 SE. Correlations are plotted separately for each condition.

Study 2

When individual cues are insufficient – eliminating some, but not all potential referents – children must rely on multiple, cooperating cues in order to successfully learn new words. We predicted that children would be able to use the conjunction of both a speaker's gaze and ME to learn and retain the names of novel objects.

Method

Participants. The final sample included 40 children (28 female, 12 male) with an average age of 40.1 months (range: 38-42 months). This is the same age range as Study 1 but did not include any children who participated in Study 1. Children were recruited using the same procedure as Study 1 with the same exclusionary criteria. The demographics of the final sample included 39 children who were Caucasian and one child who was Caucasian and Asian. One child was Hispanic/Latino. One additional child was tested but not included in the final sample, because they ended the experiment early.

Procedure. Participation involved the same procedure as Study 1 except that children completed a different version of the word learning task.

Word learning task. The same apparatus was used as in Study 1. Since the final sample consisted of children who had not participated in Study 1, the same novel words and objects were used as in Study 2. Only a subset of four novel objects and words were used from Study 1 (pairs 1 and 3 from Figure 2). Fewer novel stimuli were required, because there was only one condition.

Teaching trials. Each trial consisted of a video recording of an adult seated at a table with four objects. Two objects were positioned on the table to the left of the adult, the other two objects were positioned on the table to the right of the adult.[5] Each pair of objects on each side of the table consisted of one novel object and one familiar object with a known name. Thus, there were two novel objects and two familiar objects in total. At the beginning of each video, the adult looked into the camera, then looked towards the objects on the right side of the table, then

[5] Because the objects were super-imposed digital images, the same raw videos were used from Study 1.

towards the objects on the left side of the table, and then returned her gaze to the camera. Finally, she labeled one of the objects (e.g., “It’s a juff”). On some trials, she labeled a familiar object with a known name, while on other trials she labeled a novel object. Immediately after labeling the object, the speaker then looked towards the side of the table where the target object was located. Due to the proximity of the objects on each side of the table, it was impossible to determine which object from the pair the speaker was fixating (see Figure 10).

In order to identify the referents of novel words, children must rely on multiple cues. This is because individual cues eliminate some, but not all of the distractors. Using just ME, children could rule out the two familiar objects with known names, leaving the two novel objects as potential referents. Using just the speaker’s gaze, children could rule out the pair of objects on the other side of the table, leaving the two objects in the vicinity of the speaker’s gaze as potential referents. Children could unambiguously identify the referents of the novel words only by using both the speaker’s gaze and ME. The two cues in conjunction eliminated all distractors, leaving only the correct referent of the novel word.

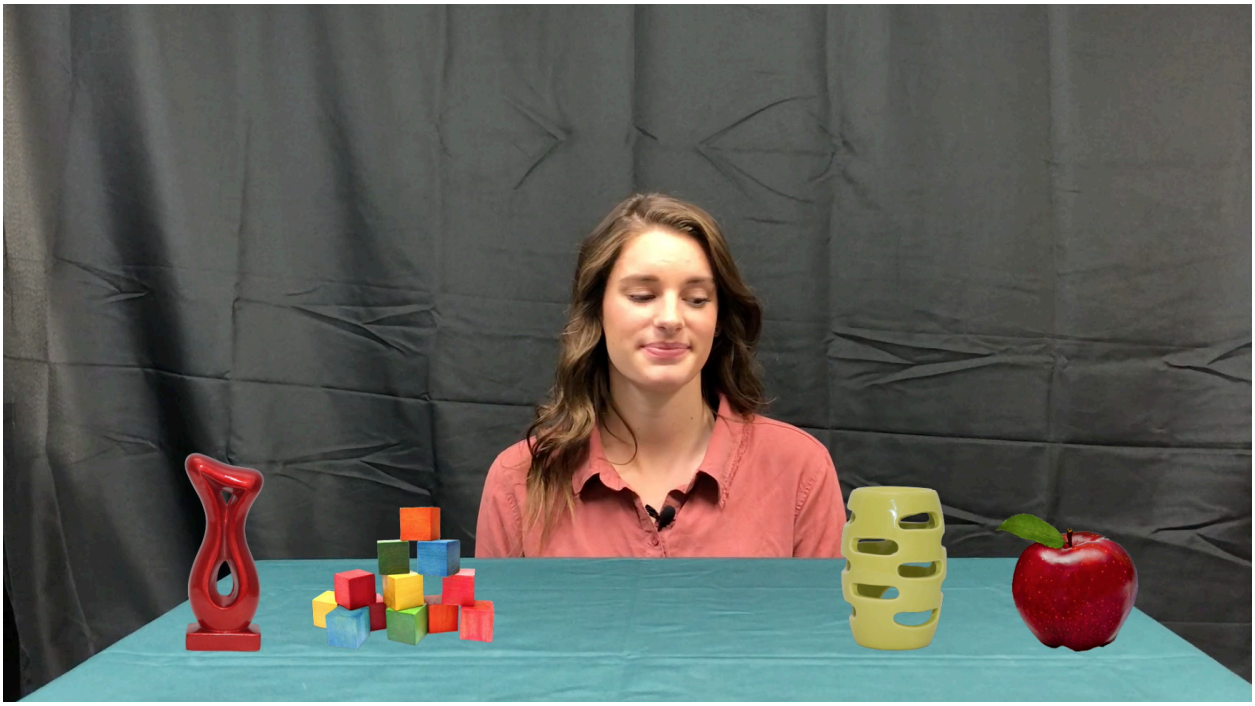

Figure 10. Still frame from a teaching video in Study 2.

Test trials. On each trial, children saw images of two objects – one image was displayed in the bottom left corner of the screen, the other image in the bottom right corner. The images were familiar and novel objects placed on gray backgrounds. Children then heard a sentence labelling one of the objects by name followed by a generic sentence to maintain their attention (e.g., "Where's the modi? That's cool!"). These sentences were produced by the same adult in the teaching videos. On some trials, both images were familiar objects with known names. On other trials, both images were novel objects from teaching trials. Images and sound files were the same as those used in Study 1.

Trial order. The entire experiment consisted of 26 trials. This included 11 teaching trials and 15 test trials. On eight teaching trials, the speaker labeled a novel object (each novel object was the target twice). On three teaching trials, the speaker labeled a familiar object with a known

name (juice, book, flower).[6] On 12 test trials, the target was a novel object (each novel object was the target three times). On three test trials, the target was a familiar object (dog, boat, cat). For test trials with novel objects, the distractor was another novel object. For test trials with familiar objects, the distractors were other familiar objects (bird, towel, horse). Each novel object occurred equally often as a target and distractor on both teaching and test trials.

Trials were arranged in a pseudo-random order. Children first saw all the teaching trials and then all the test trials. The target object was never on the same side of the screen for more than three consecutive trials. The target was never the same object for more than two consecutive trials. The location of the targets and distractors was balanced as closely as possible given the odd number of trials. On teaching trials, the target occurred six times on the left and five times on the right. On test trials, the target occurred eight times on the left and seven times on the right. These constraints were necessary to make sure that children did not notice any regularities in trial structure that they could use to fixate the target before it was labeled.

Two different trial orders were created to counterbalance the assignment of novel names to novel objects. In contrast to Study 1, different novel word-object pairings were used between children. Study 2 does not compare learning between different conditions, but rather whether children are able to learn novel words at all. This counterbalancing was therefore necessary to ensure that learning did not result from anything idiosyncratic to the stimuli (e.g., learning occurred because a certain novel object looks inherently more like a *juff* than the other novel objects).

[6] Different novel objects were used for trials with a familiar target; these novel objects were never labeled and did not appear on test trials.

Data collection. Children's eye movements were tracked using a Tobii X2-60 eye tracker. The Tobii X2-60 tracks the location (in x- and y-coordinates) of children's fixations every 16 ms. Regions of Interest (ROIs) were determined separately for teaching and test trials. For teaching trials these were: the target object, the three distractor objects, and the speaker's face. Given the close proximity of the objects on either side of the table, however, it was not possible to reliably discriminate between children's fixations to individual objects (see Figure 11). Children's fixations therefore were not analyzed for teaching videos. These trials, however, are less important in Study 2, because our primary question was not whether children could identify the referent, but rather learn and retain the novel words. This question is best addressed by analyzing children's fixations on test trials. For test trials the ROIs were the target and distractor image; the exact x- and y-coordinates for these ROIs were known, since the image locations were determined via python.

The dependent variable for all analyses is the proportion of time children spent fixating the target object out of the total time spent fixating both the target and distractor objects. Frames for which children were not fixating either object or children's fixation location could not be tracked (due to blinking, pointing, tilting of the head, etc.) were coded as NAs and treated as missing data. Proportions were calculated during a critical window 300 to 1,800 ms after the onset of the target word on each trial.

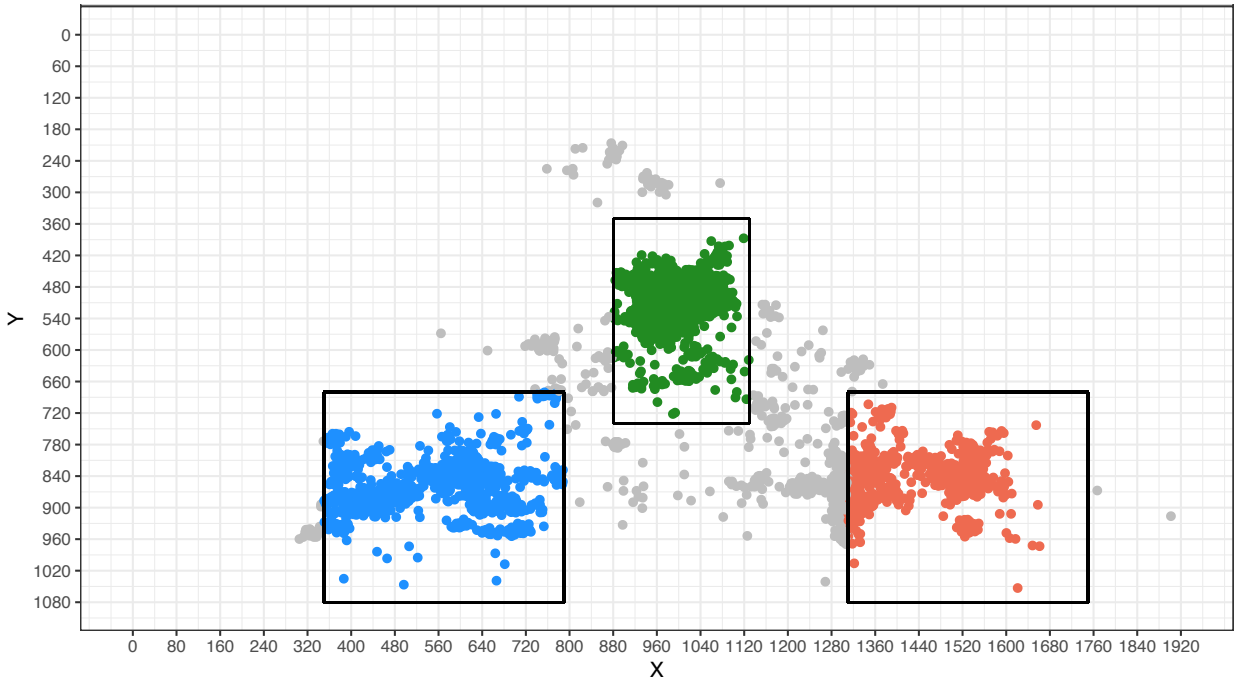

Figure 11. Gaze locations for one participant are plotted as a function of their horizontal and vertical pixel location on the television screen. Data points represent all of the fixations for all teaching trials. Data points are color coded if they fall within the ROIs for the speaker's face (green), objects on the left side of the table (blue), objects on the right side of the table (red), or elsewhere on the screen (gray). Distinct, but overlapping, clusters of fixations to each of the objects within the ROIs on either side of the table are observable.

Data cleaning. Individual trials were excluded from analyses if the child was not fixating either object for more than half (750 ms) of the critical window. These trials were excluded, because there was too much missing data. After cleaning, we identified 11 children who had fewer than two useable test trials. Video recordings for these children were hand coded. Using custom software, coders determined for each frame whether the child was looking at the left image, right image, or neither image. Coders were unaware of the target object, target location, and condition for each trial. After hand coding, no child had too much missing data. Tobii and

handcoded data were combined by downsampling the Tobii data from 60 Hz to 30 Hz by binning every 33 ms and calculating children's average fixation proportion within each bin.

Children in the final sample had on average 9.12 (SD=2.5) useable test trials (out of the maximum of 12).

Statistical Analyses. The proportion of children's fixations to the target object on test trials was centered by subtracting 0.5. This centered proportion was regressed on an intercept. The full random effects structures were included in all models (Barr, Levy, Scheepers, & Tily's, 2013). The analysis was carried out using a linear mixed effects models at the trial level that was fit using maximum likelihood estimation. Estimates of degrees of freedom and significance tests were completed using the Kenward-Roger procedure. If the proportion of children's fixations to the target object is above chance (i.e., above 0 after centering), then the inclusion of the intercept will significantly improve the model fit.

Result

Test trials. Changes in the proportion of children's fixations to the target object out of their total fixations to the target and distractor objects over the course of test trials are plotted in Figure 12. Children were equally likely to fixate the target and distractor object (i.e., 50% accuracy) at the onset of the target word. The proportion of children's fixations to the target object increased over time.

Target fixations. We predicted that children would learn and retain the novel word-object pairings from teaching trials. Therefore, their accuracy in fixating the target object would be significantly higher than chance on test trials. Children's accuracy in fixating the target object

during the critical window (M = 57.9%, SD = 14.1%) was indeed significantly higher than chance, b = 0.08, $F(1,37.1) = 14.0$, $p < .001$ (see Figure 13).

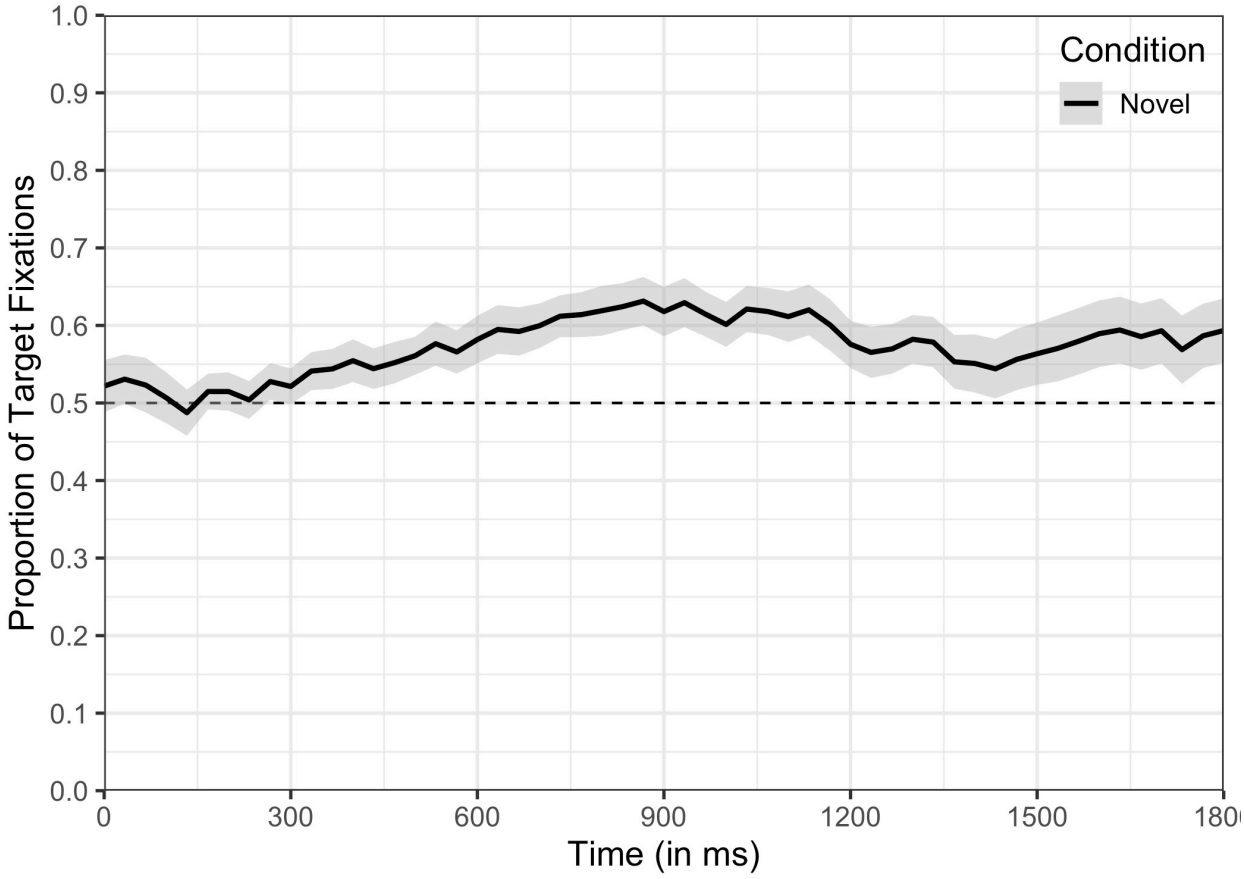

Figure 12. Time course of changes in the proportion of children's fixations to the target object out of total fixations to the target and distractor objects on test trials. Fixation proportions are plotted for trials with novel objects as a function of time (in ms) since the onset of the target word. Solid lines represent the proportion of fixations averaged across trials and participants. Ribbons around the lines represent +/- 1 SE. The horizontal dashed line in black represents chance (i.e., equal fixations to both the target and distractor object).

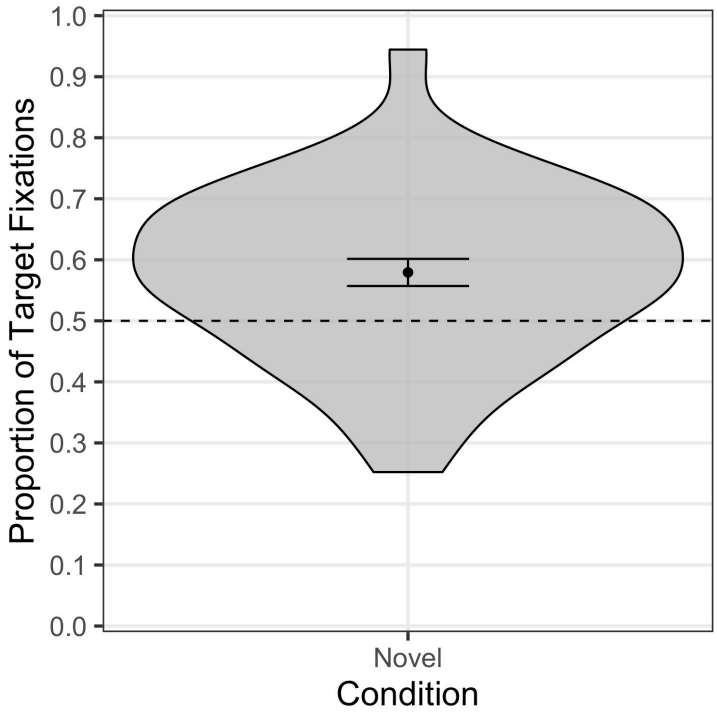

Figure 13. Proportion of children's fixations to the target object out of their total fixations to the target and distractor objects during a critical window 300 to 1,800 ms after the onset of the target word on teaching trials. Proportions are plotted only for trials with novel objects. Data points represent the average fixation proportion of all children. Error bars represent +/- 1 SE. Violins represent the distribution of fixation proportions across children (i.e., wider envelopes indicate more children with an average fixation proportion at that value than values with narrower envelopes.

General Discussion

When children first hear a novel word, there are many cues in their environment that they can use to help identify its referent. The current research examined how this abundance of information affects children's novel word learning in different ways.

In Study 1, children watched videos of an adult producing novel words. We found that children were the most accurate in identifying the referents of these novel words when they were able to use both the speaker's gaze and ME compared to when they were only able to use one of these cues. The decrease in accuracy was more pronounced when children could only rely on the speaker's gaze compared to when they could only rely on ME. On subsequent test trials, however, we found that children were equally successful in learning and retaining these novel words, regardless of the learning context. Moreover, individual differences in children's accuracy on the test trials were not correlated with their accuracy during learning. Children who were more successful in identifying the referent of the novel word were not more successful in learning and remembering its name. Thus, when a single cue will suffice, the presence of additional cues may aid children in better identifying the referent of a novel word *without* affecting word learning, at least in the current paradigm.

In Study 2, children again watched videos of an adult producing novel words. Unlike the previous study, however, individual cues were insufficient – each cue eliminated some, but not all potential referents. Children could only identify the correct referents of novel words by using both the speaker's gaze and ME. On subsequent test trials, we found that children were successful in learning and retaining these novel words. Together with Study 1, these results demonstrate that the effect of multiple cues on word learning is context dependent. When children do not need to combine cues in order to learn new words, the presence of multiple cues is superfluous and does not affect learning. When children *must* combine cues in order to learn new words, however, children can and do succeed in learning.

The following sections explore each of these findings in greater detail – addressing how the results support or contradict our hypotheses, how the results fit within the broader literature,

methodological limitations and reasons for caution when interpreting the results, and important directions for future research.

Asymmetric benefits to referent selection

The improvements in children's ability to identify the referents of novel words when using multiple cues in Study 1 are only partially consistent with our hypotheses. Consistent with our predictions, children always benefited from the addition of a second cue. We did not predict, however, that this benefit would be stronger when the addition was ME compared to the speaker's gaze. This asymmetry should be interpreted with caution, in part, because there are other ways to quantify children's attention. When children hear a novel word, what may matter the most is not how much time children spend fixating the target object out of everything they look at (i.e., target object, distractor object, speaker's face), but rather the balance in how much children attend to each object. Put another way, children likely do not consider that a novel word like *juff* labels the speaker's face. Instead, children are trying to determine which of the two objects on the table is the *juff*. Indeed, when children's fixations are instead quantified as the proportion of time spent fixating the target object out of the total time spent fixating either the target or the distractor object, there are no significant differences between Gaze (M = 66.6%, SD = 22.0%), ME (M = 65.0%, SD = 19.8%), and Gaze+ME (M = 71.2%, SD = 16.7%) teaching trials, $F(2, 46.6) = 0.98$, $p = 0.38$. After hearing the novel word, children spend more time fixating the intended referent, rather than the unintended referent. This imbalance in attention is the same regardless of the number of cues that are available to children.

Other reasons to be cautious in interpreting the asymmetry between speaker gaze and ME are potential limitations from the experimental design. The use of video recordings enabled

precision in the timing of stimulus presentation, counterbalancing between conditions, and facilitated tracking children's eye movements (although it is possible to track children's eye movements during live interactions, e.g., Yu & Smith, 2013). Social cues, like a speaker's gaze, however, may be fundamentally different in video recordings compared to live interactions. For instance, the speaker in our videos did not interact with children or respond contingently to their behaviors as they would in a live interaction. Video recordings may therefore limit how much children attend to a speaker's gaze and lead to underestimates of its reliability as a cue in the current studies.

Despite these methodological concerns, the asymmetries we observed between speaker gaze and ME are consistent with prior research that involved live interactions. Children of a similar age to our participants struggle to use just a speaker's gaze to learn new words (Booth, McGregor & Rohlfing, 2008) and preferentially use ME over just a speaker's gaze to learn new words (Jaswal & Hansen, 2006). Moreover, parents and children coordinate their visual attention to objects not by tracking one another's eye movements, but rather hand movements (Yu & Smith, 2013).

While the results from the current studies are suggestive, it is important for future research to more systematically examine potential asymmetries in how combinations of different cues affect children's ability to identify the referents of novel words. This work should involve further manipulations to the cues used in the current experiment. For instance, testing whether the asymmetry between gaze and ME can be reduced or flipped by increasing the strength of a speaker's gaze (e.g., making changes in a speaker's gaze more prominent) and decreasing the strength of ME (e.g., by using familiar objects whose labels are less well-known by children). Speaker gaze and ME, however, are not the only cues children use to identify the referents of

novel words; further research should examine whether asymmetries exist between different word learning cues. Finally, more work is necessary to determine whether and how the benefit children receive from the presence of multiple word learning cues is affected by changes in the amount of exposure children have and changes with development.

Equal success in word learning

Contrary to our hypothesis, the degree to which children succeeded in learning new words in Study 1 was unaffected by the presence of multiple, cooperating cues. Despite being more accurate in identifying the referent of novel objects when they were able to use two cues, children were no more accurate in learning and retaining these word-object pairings than those trained with a single cue. Null results, however, should always be interpreted with caution. Given the significant amount of variability between children (mean accuracy on test trials was 61.2% with a standard deviation of 23.4%), it is possible that true differences in word learning did exist between the conditions, but we are unable to detect these differences statistically (particularly if they are subtle). In fact, many studies of novel word learning that find above chance performance in one condition and at chance performance in another condition, nevertheless fail to find significant differences between conditions (e.g., Pomper & Saffran, 2018). Increasing the amount of data (i.e., test trials) per child would help overcome this limitation. Such changes, however, are often not feasible for word learning studies. Children can only be taught so many novel words in a single experiment and can only be tested so many times on the same novel word before memory and attentional demands become too great. Indeed, in Study 1 there were significant reductions in the amount of useable data by the end of the experiment.

Alternatively, it is possible that these null results are accurate – that there is indeed no difference between children's word learning when they are able to use two cues compared to one cue. In support of this possibility, prior research has found that improvements in children's accuracy during referent selection do not always translate into improvements in learning. For instance, supplementing the speaker's gaze with social cues (like the speaker holding the novel object) increases children's accuracy in attending to a novel object when it is labeled, but does not improve their success in learning and retaining the novel object's label (Booth, McGregor, & Rohlfing, 2008). Similarly, manipulations that do not affect children's accuracy in referent selection may nevertheless affect word learning. For instance, decreasing the number of familiar objects or increasing how often familiar objects are repeatedly used does not affect children's ability to identify the referent of a novel word using ME, but does improve children's accuracy in learning and retaining the novel word (Axelsson & Horst, 2014; Horst, Scott, & Pollard, 2010). Together, these results suggest that there may be factors that independently affect referent selection and word learning. Simply increasing the amount of time children spend looking at a novel object when it is labeled does not guarantee that children will be more successful in learning.

That there may in fact be no differences between children's word learning as a function of the manipulations in Study 1 does not mean that this will always be the case. It is possible that the addition of a second cue would have benefited word learning under different circumstances. For instance, the benefit to word learning from a second cue may only occur with certain amounts of exposure. This can be best exemplified at the extremes. With too little exposure, children will fail to learn new words; with too much exposure, children will perfectly learn all words. When children are performing to close to floor or ceiling, the addition of a second cue

may fail to affect word learning. Similarly, the effect of additional cues on word learning may vary based on children's age. In Study 1, individual cues were sufficient for 3-year-olds to learn new words. At younger ages, however, children fail to learn new words using just speaker gaze or just ME. For these children, the addition of a second cue may not be superfluous, but necessary for word learning (in a way analogous to Study 2). For these reasons, it is important for future research to further explore how the presence of both the speaker's gaze and ME affect word learning across different levels of exposure and across children's development.

Uncorrelated referent selection and word learning

Contrary to our hypothesis, individual differences in children's accuracy during referent selection were not correlated with individual differences in children's accuracy in learning. The same challenges highlighted earlier apply in interpreting this null result, including potential methodological limitations in noise due to insufficient data. There have been mixed results, however, when examining correlations between teaching and test accuracy in the novel word learning literature. Research involving wider age ranges of children has found significant correlations between children's accuracy in referent selection and their accuracy at test (Bion, Borovsky, & Fernald, 2013). However, other research using a narrower age range (one that is identical to the current studies) failed to find a correlation between children's accuracy in referent selection and word learning (Pomper & Saffran, 2018). One possibility is that wider age ranges yield significant correlations because there is greater variability in accuracy during both referent selection and word learning. Considering the substantial amount of variability (in both referent selection and word learning) even within the narrow age range of the current experiment and that improvements in children's accuracy in referent selection as a group (i.e., between

conditions) did not yield improvements in word learning as a group in Study 1, it is possible that there truly is no correlation between children's accuracy during referent selection and their success in word learning.

Multiple, imperfect cues

In Study 2, individual cues were no longer sufficient. In order to correctly identify the referents of novel words, children needed to use multiple cues during teaching videos. Consistent with our hypothesis, children's accuracy in identifying the target novel object was significantly higher than chance on subsequent test trials. Although simplistic, these results are perhaps the most important. With only the results from Study 1, we might be inclined to conclude that multiple cues are unnecessary for word learning. The results from Study 2, however, demonstrate that multiple cues may be crucially important, because they enable word learning that would otherwise be impossible. Probabilistic cue combination is a phenomenon that is not limited to language but extends to other cognitive domains and biology more broadly. It is something that humans are remarkably good at doing and may even be a domain in which infants and children outperform adults (e.g., Yurovsky, Boyer, Smith, & Yu, 2013). In contrast to our lab environments, children are learning words in an imperfect world. While research involving individual cues is important, these experiments may be lacking important external validity. A better understanding of how children learn new words, therefore, requires that future work further examine what cues are present in naturalistic environments and the reliability and accuracy of these cues in identifying the referents of words. Moreover, future work must examine how children's ability to combine multiple, imperfect cues in order to learn new words

may vary between environments (when the number and types of cues vary) and across development.

Broader impacts

The current studies represent an important first step in understanding how children combine cues to learn new words. There is more research to be done further exploring how typically-developing children use multiple cues, including some of the future directions discussed earlier. One of the most important future directions for this work, however, is expanding the sample of participants to include children who struggle to learn language, like children with autism and children with developmental language disorder (DLD).

A prominent hypothesis in the autism literature is that children with autism may struggle to learn new words because they are unable to use social cues like a speaker's gaze. There have been, however, mixed results in support of this hypothesis. Some research has found that children with autism are worse than their typically-developing peers in using speaker gaze to learn new words (Akechi, Kikuchi, Tojo, Osanai, & Hasegawa, 2011; Baron-Cohen, Baldwin, & Crowson, 1997; Gliga, Elsabbagh, Johson, Hudry, & Charman, 2012; Parrish-Morris, Hennon, Hirsh-Pasek, Golinkoff, & Tager-Flusberg, 2007; Preissler & Carey, 2005), while other research has found that children with autism are equally successful as their typically-developing peers (Luyster & Lord, 2009; Patrick, Hurewitz, & Booth, 2013). When a speaker's gaze is the only cue available – the only show in town – then children with autism may be inclined to use gaze. More reliable differences in word learning, however, may emerge between children with autism and their typically-developing peers when there are multiple cues available and children can choose to attend to or ignore gaze.

Research involving both children and adults with DLD has consistently found deficits in word learning. These deficits do not appear to arise from a difficulty in mapping words to objects, but rather encoding phonological information, the sounds of the novel words (McGregor, Arbisi-Kelm, Eden, & Oleson, 2020; McGregor, Gordon, Eden, Arbisi-Kelm, & Oleson, 2020). The majority of this research, however, has involved ostensive teaching – when children and adults hear novel words there was no ambiguity, because only a single referent was present. It remains unclear whether and how children and adults with DLD use individual cues to learn new words, much less how they combine multiple cues when learning.

A better understanding of how children learn new words and how this learning may go awry for children who struggle to learn language will help inform interventions that seek to improve language outcomes. In order to improve outcomes, we must understand which interventions are efficacious and what aspects of language should be targeted (e.g., Rogers & Vismara, 2008).

Conclusion

In the current era of big data, where a plethora of data is available (including years of video and audio recordings of a child's every waking moment; Roy, Frank, DeCamp, Miller, & Roy, 2015), research is not limited by a lack of data, but rather the ability to process the data. There is a growing appreciation within language research that children face a similar problem when learning language. Debates are no longer focused on whether there is enough information in children's environments to learn language, but instead on how much of this information children are able to process (e.g., Smith, Suanda, & Yu, 2014; Trueswell, Medina, Hafri, & Gleitman, 2013). The current studies represent a further step in this direction, taking for granted

the wealth of information that is available to children and attempting to better understand how children sift through this information. This research, hopefully, represents an important first step in exploring how the presence of multiple, cooperating cues affects children's ability to learn new words.

References

Akechi, H., Senju, A., Kikuchi, Y., Tojo, Y., Osanai, H., & Hasegawa, T. (2011). Do children with ASD use referential gaze to learn the name of an object? An eye- tracking study. *Research in Autism Spectrum Disorders, 5*, 1230-1242.

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning words through overhearing. *Child Development*, *72*(2), 416-430.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257-262

Axelsson, E. L., Churchley, K., & Horst, J. S. (2012). The right thing at the right time: why ostensive naming facilitates word learning. *Frontiers in Psychology*, *3*, 1-8.

Axelsson, E. L., & Horst, J. S. (2014). Contextual repetition facilitates word learning via fast mapping. *Acta Psychologica*, *152*, 95-99

Bååth, R. (2010). ChildFreq: An Online Tool to Explore Word Frequencies in Child Language. *LUCS Minor*, 16.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*(5), 874-890.

Baldwin, D. A. (1993a). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, *20*(2), 395-418.

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

Baldwin, D. A., & Tomasello, M. (1998). Word learning: A window on early pragmatic understanding. In E. V. Clark (Ed.), *The proceedings of the twenty-ninth annual child language research forum* (pp. 3-23). Chicago, IL, US: Center for the Study of Language

and Information.

Baron-Cohen, S., Baldwin, D. A., & Crowson, M. (1997). Do children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development, 68*(1), 48-57.

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

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT press.

Booth, A. E., McGregor, K. K., & Rohlfing, K. J. (2008). Socio-pragmatics and attention: Contributions to gesturally guided word learning in toddlers. *Language Learning and Development*, *4*(3), 179-202.

Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: how infants interpret adult looking behavior. *Developmental Psychology*, *38*(6), 958.

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, *8*(6), 535-543.

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*(1), 207-220.

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. *Developmental Science*, *12*(5), 815-823.

Byers-Heinlein, K., & Werker, J. F. (2013). Lexicon structure and the disambiguation of novel words: Evidence from bilingual infants. *Cognition*, *128*(3), 407-416.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17-29.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i-174.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, *13*(2-3), 221-268.

Clark, E. V. (1983). Meanings and concepts. In J. H. Flavell & E. M. Markman (Eds.), *Handbook of child psychology: Vol. III. Cognitive Development.* New York: John Wiley & Sons.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*(5), 630-641.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody picture vocabulary test: Manual*. Pearson.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763-13768.

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., Weinbergy, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228-231.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. In I.A. Sekerina, E.M. Fernandez, &

H. Clahsen (Eds.) *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97-135). Philadelphia, PA: John Benjamins Publishing Company.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, *44*(3), 677-694.

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.

Gangopadhyay, I. & Kaushanskaya, M. (2020). The role of speaker gaze and mutual exclusivity in novel word learning by monolingual and bilingual children. *Journal of Experimental Child Psychology, 197*, 104878.

Gliga, T., Elsabbagh, M., Hudry, K., Charman, T., Johnson, M. H., & The BASIS Team (2012). Gaze following, gaze reading, and word learning in children at risk for autism. *Child Development, 83*(2), 926-938.

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.

Golinkoff, R. M., Hirsh-Pasek, K., Baduini, C., & Lavallee, A. (1985, October). What's in a word? The young child's predisposition to use lexical contrast. In *Boston University Conference on Child Language, Boston*.

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: the case for a developmental lexical principles framework. *Journal of Child Language*, *21*(1), 125-155.

Golinkoff, R. M., Shuff-Bailey, M., Olguin, R., & Ruan, W. (1995). Young children extend

novel words at the basic level: Evidence for the principle of categorical scope. *Developmental Psychology*, *31*(3), 494-507.

Grassmann, S., Schulze, C., & Tomasello, M. (2015). Children's level of word knowledge predicts their exclusion of familiar objects as referents of novel words. *Frontiers in Psychology*, *6*(1200).

Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science*, *13*(1), 252-263.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23-B34.

Harris, M., Jones, D., & Grant, J. (1983). The nonverbal context of mothers' speech to infants. *First Language*, *4*(10), 21-30.

Hirsh-Pasek, K., & Golinkoff, R. M. (2008). King solomon's take on word learning: An integrative account from the radical middle. In R. V. Kail (Eds.) *Advances in Child Development and Behavior* (Vol. 36, pp. 1-29). JAI.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R.M., Brand, R.J., Brown, E., Chung, H.L., Hennon, E., Rocroi, C., & Bloom, L. (2000). Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning. *Monographs of the Society for Research in Child Development, 65*(3), 1-135.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393-1409.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128-157.

Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's new? Children

prefer novelty in referent selection. *Cognition*, *118*(2), 234-244.

Horst, J. S., Scott, E. J., & Pollard, J. A. (2010). The role of competition in word learning via referent selection. *Developmental Science*, *13*(5), 706-713.

Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy*, *15*(2), 125-150.

Hung, W. Y., Patrycia, F., & Yow, W. Q. (2015). Bilingual children weigh speaker's referential cues and word-learning heuristics differently in different language contexts when interpreting a speaker's intent. *Frontiers in Psychology*, *6*(796).

Hutchinson, J. E. (1986). Children's sensitivity to the contrastive use of object category terms. *Papers and Reports on Child Language Development*, *25*, 49-55.

Jaswal, V. K., & Hansen, M. B. (2006). Learning words: Children disregard some pragmatic information that conflicts with mutual exclusivity. *Developmental Science*, *9*(2), 158-165.

Law, F., & Edwards, J. R. (2015). Effects of vocabulary size on online lexical processing by preschoolers. *Language Learning and Development*, *11*(4), 331-355.

Luyster, R., & Lord, C. (2009). Word learning in children with autism spectrum disorders. *Developmental Psychology, 45*(6), 1774-1786.

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

Markman, E. M. (1987). How children constrain the possible meanings of words. In U. Neisser (Ed.), *Emory symposia in cognition, 1. Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 255-287). New York, NY, US: Cambridge University Press.

Markman, E. M. (1992). Constraints on word learning: Speculations about their nature, origins,

and domain specificity. In M. R. Gunnar & M. Maratsos (Eds.), *The Minnesota symposia on child psychology, Vol. 25. Modularity and constraints in language and cognition* (pp. 59-101). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology*, *16*(1), 1-27.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*, 121-157.

Masur, E. F. (1997). Maternal labelling of novel and familiar objects: Implications for children's development of lexical constraints. *Journal of Child Language*, *24*(2), 427-439.

Mather, E., & Plunkett, K. (2010). Novel labels support 10-month-olds' attention to novel objects. *Journal of Experimental Child Psychology*, *105*(3), 232-242.

McGregor, K. K., Arbisi-Kelm, T., Eden, N., & Oleson, J. (2020). The word learning profile of adults with developmenetal language disorder. *Autism and Developmental Language Impairments,* 5, 1-19.

McGregor, K. K., Gordon, K., Eden, N., Arbisi-Kelm, T., & Oleson, J. (2017). Encoding deficits impede word learning and memory in adults with developmental language disorders. *Journal of Speech, Language, and Hearing Research, 60*, 2891-2905.

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, i-129.

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name–nameless category (N3C) principle. *Child Development*, *65*(6), 1646-1662.

Messer, J. (1978). The integration of mothers' referential speech with joint play. *Child*

Development, 49, 781–787.

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science*, *9*(1), 21-34.

Moore, C., Angelopoulos, M., & Bennett, P. (1999). Word learning in the context of referential and salience cues. *Developmental Psychology*, *35*(1), 60-68.

Patrick, K. E., Hurewitz, F., & Booth, A. E. (2013). Word-mapping in autism: Evidence for backwards bootstrapping of social gaze strategies. In *Boston University Conference on Language Development Proceedings*. Cascadilla Press.

Parish-Morris, J., Hennon, E. A., Hirsh-Pasek, K., Golinkoff, R. M., & Tager-Flusberg, H. (2007). Children with autism illuminate the role of social intention in word learning. *Child Development, 78*(4), 1265-1287.

Pomper, R., & Saffran, J. R. (2018). Familiar Object Salience Affects Novel Word Learning. *Child Development*.

Preissler, M. A., & Carey, S. (2005). The role of inferences about referential intent in word learning: Evidence from autism. *Cognition, 97*, B13-B23.

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*(2), 266-280.

Rogers, S. J., & Vismara, L. A. (2008). Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child and Adolescent Psychology, 37*(1), 8-38.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. Predicting the birth of a spoken word. *Proceedings in the National Academy of Sciences, 112*(41), 12663-12668.

Samuelson, L. K., & Smith, L. B. (1998). Memory and attention make smart word learning: An alternative account of Akhtar, Carpenter, and Tomasello. *Child Development*, *69*(1), 94-104.

Samuelson, L., & Smith, L. B. (2000). Grounding development in cognitive processes. *Child Development*, *71*(1), 98-106.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word–referent learning. *Trends in Cognitive Sciences*, *18*(5), 251-258.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568.

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: Individual differences in statistical word-referent learning in infants. *Language Learning and Development, 9*(1), 25-49.

Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition*, *38*(2), 179-211.

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development*, *72*(3), 748-767.

Tomasello, M. (1999). Having intentions, understanding intentions, and understanding communicative intentions. In P. D. Zelazo, J. W. Astington, & D. R. Olson (Eds.), *Developing theories of intention: Social understanding and self-control* (pp. 63-75). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, *10*(4), 401-413.

Tomasello, M., & Barton, M. E. (1994). Learning words in nonostensive

Contexts, *Developmental Psychology*, *30*(5), 639-650.

Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language*, *23*(1), 157-176

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126-156.

Vincent-Smith, L., Bricker, D., & Bricker, W. (1974). Acquisition of receptive vocabulary in the toddler-age child. *Child Development*, 189-193.

Waxman, S., & Gelman, R. (1986). Preschoolers' use of superordinate relations in classification and language. *Cognitive Development*, *1*(2), 139-156.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1-34.

Woodward, A. L. (2000). Constraining the problem space in early word learning. In R. M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L.B. Smith, A.L. Woodward, N. Akhtar, M. Tomasello, & G. Hollich (Eds.) *Becoming a word learner: A debate on lexical acquisition* (pp. 81-114). Oxford University Press.

Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science*, *6*(3), 297-311.

Woodward, A. L. (2004). Infants' Use of Action Knowledge to Get a Grasp. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a Lexicon* (pp. 149-171). Cambridge, MA, US: MIT Press.

Woodward, A. L., & Markman, E. M. (1991) Constraints on Learning as Default Assumptions: Comments on Merriman and Bowman's "The Mutual Exclusivity Bias in Children's Word Learning." *Developmental Review*, *11*(2), 137-163.

Xu, Y., Regier, T., & Newcombe, N. S. (2017). An adaptive cue combination model of human

spatial reorientation. *Cognition*, *163*, 56-66.

Yow, W. Q., and Hung, W. Y. (2013). Impact of bilingual (code- switching) experience on preschoolers? Sensitivity to pragmatic cues. *Paper Presented at the Society of Research in Child Development*, Seattle.

Yow, W. Q., Li, X., Lam, S., Gliga, T., Chong, Y. S., Kwek, K., & Broekman, B. F. (2017). A bilingual advantage in 54-month-olds' use of referential cues in fast mapping. *Developmental Science*, *20*(1).

Yow, W. Q., & Markman, E. M. (2011). Young bilingual children's heightened sensitivity to referential cues. *Journal of Cognition and Development*, *12*(1), 12-31.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, *8*(11), e79659.

Yurovsky, D., Boyer, T. W., Smith, L. B., & Yu, C. (2013). Probabilistic cue combination: Less is more. *Developmental Science, 16*(2), 149-158.

Yurovsky, D., & Frank, M. C. (2017). Beyond naïve cue combination: Salience and social cues in early word learning. *Developmental Science*, *20*(2).

Yurovsky, D., Wade, A., & Frank, M. (2013). Online processing of speech and social information in early word learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35), 1641-1646.

Appendix

Tr.Num	Sound Stimulus	Left Image	Right Image	Target Side	Block	Tr.Type	Condition
1	Juice	novel7	juice	R	1	Teaching	Familiar
2	Tever_Its_Gaze	novel1	novel2	L	1	Teaching	Gaze
3	Juff_See_Gaze	novel1	novel2	R	1	Teaching	Gaze
4	Juff_That_Gaze	novel2	novel1	L	1	Teaching	Gaze
5	Tever_Look_Gaze	novel2	novel1	R	1	Teaching	Gaze
6	Dog_Wow	dog	bird	L	1	Teaching	Familiar
7	Tever_Find_Check	novel1	novel2	L	1	Test	Gaze
8	Juff_Look_Wow	novel2	novel1	L	1	Test	Gaze
9	Tever_Where_Check	novel2	novel1	R	1	Test	Gaze
10	Juff_Look_Cool	novel1	novel2	R	1	Test	Gaze
11	Juff_Where_Wow	novel2	novel1	L	1	Test	Gaze
12	Tever_Find_Cool	novel2	novel1	R	1	Test	Gaze
13	Book	book	novel8	L	2	Teaching	Familiar
14	Blicket_Its_ME	novel4	ball	L	2	Teaching	ME
15	Rel_That_ME	cheese	novel3	R	2	Teaching	ME
16	Blicket_Look_ME	apple	novel4	R	2	Teaching	ME
17	Rel_See_ME	novel3	bear	L	2	Teaching	ME
18	Boat_Cool	boat	towel	L	2	Teaching	Familiar
19	Rel_Look_Check	novel4	novel3	R	2	Test	ME
20	Blicket_Where_Wow	novel3	novel4	R	2	Test	ME
21	Rel_Where_Cool	novel4	novel3	R	2	Test	ME
22	Rel_Find_Check	novel3	novel4	L	2	Test	ME
23	Blicket_Find_Cool	novel3	novel4	R	2	Test	ME
24	Blicket_Look_Wow	novel4	novel3	L	2	Test	ME
25	Flower	flower	novel9	L	3	Teaching	Familiar
26	Manu_Look_GazeME	novel5	cookie	L	3	Teaching	Gaze+ME
27	Gip_See_GazeME	novel6	truck	L	3	Teaching	Gaze+ME
28	Manu_That_GazeME	duck	novel5	R	3	Teaching	Gaze+ME
29	Gip_Its_GazeME	cake	novel6	R	3	Teaching	Gaze+ME
30	Kitty_Check	horse	cat	R	3	Teaching	Familiar
31	Gip_Where_Wow	novel6	novel5	L	3	Test	Gaze+ME
32	Manu_Look_Cool	novel5	novel6	L	3	Test	Gaze+ME
33	Gip_Look_Check	novel5	novel6	R	3	Test	Gaze+ME
34	Manu_Where_Cool	novel6	novel5	R	3	Test	Gaze+ME
35	Manu_Look_Wow	novel5	novel6	L	3	Test	Gaze+ME
36	Gip_Find_Check	novel6	novel5	L	3	Test	Gaze+ME

Table 1. Example Trial Order for Study 1