Developmental change and experience in the mutual exclusivity effect

Molly Lewis[1, 6], Veronica Cristiano[2], Brenden Lake[3], Tammy Kwan[4], & Michael C. Frank[5]

[1] University of Chicago

[2] Gallaudet University

[3] New York University

[4] Cognitive Toybox, Inc.

[5] Stanford University

[6] University of Wisconsin-Madison

Developmental change and experience in the mutual exclusivity effect

Introduction

A key property of language is that words tend to have distinct meanings, and concepts tend to be referred to via unique words (Bolinger, 1977; Clark, 1987). Like a whole host of other regularities in language – for example, the existence of abstract syntactic categories – children cannot directly observe the tendency for one-to-one word-concept mapping, yet even very young children behave in a way that is consistent with it. Evidence that children obey the one-to-one regularity comes from what is known as the "mutual exclusivity" (ME) effect. In a typical demonstration of this effect (Markman & Wachtel, 1988), children are presented with a novel and familiar object (e.g., a whisk and a ball), and are asked to identify the referent of a novel word ("Show me the dax"). Children across a wide range of ages, experimental paradigms, and populations tend to choose the novel object as the referent in this task (Bion, Borovsky, & Fernald, 2013; Golinkoff, Mervis, Hirsh-Pasek, & others, 1994; Halberda, 2003; Markman, Wasow, & Hansen, 2003; Mervis, Golinkoff, & Bertrand, 1994). The goal of this paper is to review and synthesize evidence for an aspect of the mutual exclusivity behavior that has received relatively little attention in the literature: the role of experience and development.

Before engaging with the prior literature related this behavior, it is useful to first make several theoretical distinctions and clarify terminimology. Markman and Wachtel (1988)'s seminal paper coined the term "mutual exclusivity," which was meant to label the theoretical proposal that "children constrain word meanings by assuming at first that words are mutually exclusive – that each object will have one and only one label." (Markman, 1990, p. 66). That initial paper also adopted a task used by a variety of previous authors (including Golinkoff, Hirsh-Pasek, Baduini, & Lavallee, 1985; Hutchinson, 1986; Vincent-Smith, Bricker, & Bricker, 1974), in which a novel and a familiar object were presented to children in a pair and the child was asked to "show me the *x*," where *x* was a novel label. Since then, informal discussions have used the same name for the paradigm (this precise experiment), inference (the ability to disambiguate the novel word), and the effect (the fact that children select the novel object as the referent). Further, the same name is also often used as a tag for a particular theoretical account (an early assumption or bias regarding the one-to-one nature of the lexicon). This conflation of paradigm/effect with theory is problematic, as authors who have argued against the specific theoretical account then are in the awkward position of rejecting the name for the paradigm they themselves have used. Other labels (e.g. "disambiguation" or "referent selection" effect) are not ideal, however because they are not as specific and do not refer as closely to the previous literature.

ME has also been referred to as "fast mapping." We believe that this label is confusing because it conflates two distinct ideas. In an early study, Carey and Bartlett (1978) presented children with an incidental word learning scenario by using a novel color term to refer to an object: "You see those two trays over there. Bring me the *chromium* one. Not the red one, the *chromium* one." Those data (and subsequent replications, e.g. Markson & Bloom, 1997) showed that this exposure was enough to establish some representation of the link between phonological form and meaning that endured over an extended period; a subsequent clarification of this theoretical claim emphasized that these initial meanings are partial (Carey, 2010). Importantly, however, demonstrations of retention relied on learning in a case where there was a contrastive presentation of the word with a larger set of contrastive cues (Carey & Bartlett, 1978) or pre-exposure to the object (Markson & Bloom, 1997).

Further, the "fast mapping" label has been the focus of critique due to findings by Horst and Samuelson (2008) that young children do not always retain the mappings that result from the disambiguation inference. In this work, children were prestend with a novel word and asked to identify the referent in the ME paradigm, and they generally succeeded in making the correct inference (selecting the novel object). However, when asked to recall the referent of the same label after a short 5-min delay, children performed poorly. This pattern of results suggests an important distinction between making the ME inference in the context of the ME paradigm, and actually learning the meaning of the novel word such that it can be recalled later beyond the context of the ME paradigm.

Here we adopt the label "mutual exclusivity" (ME) effect as a generic term refering to the empirical finding that young children tend to map a novel word to a novel object. We distinguish the ME effect from the family of experimental paradigms that demonstrate the effect, which we refer to as "ME paradigms." Further, we distinguish the paradigm and the associated effect from the cognitive inference made by children that leads to the ME effect ("ME inference"). Each of these are in turn distinguished from theories which seek to explain the ME inference ("ME theory"). In each of these cases, we use the term "mutual exclusivity" as convenient nomenclature but do so *without* prejudgement of the theoretical account.

The ME effect has received much attention in the word learning literature because the ability to identify the meaning of a word in ambiguous contexts is, in essence, the core problem of word learning. That is, given any referential context, the meaning of a word is underdetermined (Quine, 1960), and the challenge for the world learner is to identify the referent of the word within this ambiguous context. For example, suppose a child hears the novel word "kumquat" while in the produce aisle of the grocery store. There are an infinite number of possible meanings of this word given this referential context, but the ability to make a disambiguation inference would lead her to rule out all meanings for which she already had a name. With this restricted space of possibilities, she is likely to identify the correct referent than if all objects in the context were considered as possible referents.

Being able to make an ME inference could also help children acquire words for multiple words that can be used to refer to the same object in the world, even though they actually refer to different concepts (for example, property names and object parts such as "turquoise" and "handle"; Markman & Wachtel, 1988). Consider a child who hears the novel word "turquoise" in the context of a turquoise-colored ball. If she obeys the one-to-one property of language and already knows the word "ball," the child may assume that "turquoise" refers to a property of the ball, such as color, rather than the ball itself. Of course, seeing evidence about the meaning of "turquoise" across multiple different turqouise referents situations (referred to as "cross-situational evidence"; Yu & Smith, 2007).

Making ME inferences could be particularly useful for learning subordinate (e.g., "dalmation") and superordinate labels (e.g., "animal"); Each instance of these labels is always consistent with concepts at all levels of the conceptual hierarchy (an observed dalmation is equally consistent with the labels "dalmation," "dog" and "animal"; e.g., Waxman & Gelman, 1986). Also, and unlike in the case of property words, a child will never observe cross-situational evidence that disambiguates among candidate concepts at different levels of the hierarchy. Thus ME inferences provide one possible route through which children might resolve this inherent ambiguity in word learning.

Despite – or perhaps due to – the attention that the ME effect (and the related consequences of making ME inferences) has received, there is little consensus regarding the cognitive mechanisms underlying it. Does it stem from a basic inductive bias on children's learning abilities ("constraint and bias accounts," "probabilistic accounts," and "logical inference accounts"), a learned regularity about the structure oflanguage ("overhypothesis accounts"), reasoning about the goals of communication in context ("pragmatic accounts"), or perhaps some mixture of these? Across the literature, researchers have tested a variety of populations of children and used a wide range of different paradigms in order to discriminate between these theories, and a successful theory of ME will need to be able to account for this wide range of empirical phenomena.

In the current paper, our goal is to present evidence for one particular pattern of findings related to ME that has played a relatively minor role in theorizing about ME: Developmental change in the magnitude of the effect. Characterizing developmental change is important because it provides a key constraint on theoretical accounts of ME. Namely, change in the magnitude of the ME effect must be due either to maturational change or the child's increasing experience with the world, or both. In our work here, we focus on characterizing the link between developmental change and one type of experience – linguistic experience. There are a variety of ways that linguistic experience could support the ME inference. For example, a child who knows more words in general might be more likely to know the familiar word in the ME task, and therefore more likely to select the novel object. Alternatively, linguistic experience might allow children to learn generalities about how language is used that could be helpful in making the ME inference, such as general pragmatic reasoning or an understanding of the one-to-one regulartity in language. Our aim here is not to definitively discriminate between theories of ME, but rather present evidence for a theory-constraining pattern of data. In the General Discussion, we consider in more detail how existing theories of ME might account for our findings.

Across the literature on ME, the primary focus of theorizing has been on accounting for why children at one or a few timepoints in development behave in a way that is consistent or not with the ME effect, rather than for development change in the strength of the bias. In part, this focus may be due to methodological challenges in conducting developmental experiments, rather than to an underlying theoretical motivation. Particularly, since data collection from young children is expensive, it is costly for researchers to collect data from children across more than a couple ages groups. In addition, experimental evidence from the ME paradigm is typically summarized as a binary description (children's "success" or "failure" in the ME task) rather than as a more continuous estimate of the effect, and this methodological choice may obscure evidence for more subtle changes in the cognitive system across development.

There are, however, a handful of studies that show developmental change in the mutual exclusivity effect by testing multiple age groups within the same experiment (Bion et al., 2013; e.g., Halberda, 2003; Merriman, 1986). For example, Halberda (2003) tests 14- 16- and 17- mo in the ME paradigm, and finds a pattern of developmental change: 14 mo children are biased to select the familiar object, 16-mo were at chance, and 17-mo were biased to select the novel object, demonstrating the ME effect. While multi-age-group studies such as this provide clear evidence *that* there is developmental change, they do not provide they type of quantitative description of its developmental trajectory of the effect that could provide an important constraint on theories of ME.

The Current Study

We first describe the state of the evidence for developmental change in the ME effect via a meta-analysis of the extant empirical literature. By aggregating across studies that each test different ages, the meta-analytic approach allows us to take advantage of the large number of studies already conducted on the ME effect in order to characterize developmental change. We then present two new, relatively large-sample developmental experiments that investigate the causal role of linguistic experience in contibuting to the ME effect. In Experiment 1, we examine the relationship between vocabulary size and the strength of the ME effect on a large sample of children. We find evidence that children with larger vocabularies tend to show a stronger ME effect, consistent with the notion that language experience influences the ME effect. In Experiment 2, we more directly test the hypothesis that language experience plays a *causal* role in the ME effect, by directly manipulating children's amount of experience with a word. We find greater experience with the familiar word in the ME paradigm leads to a stronger ME effect. In the General Discussion, we conclude by discussing the role of developmental change and experience in the context of candidate theories of ME, in the context of our evidence.

Meta-analysis

To assess the strength of the disambiguation bias as well a moderating factors, we conducted a meta-analysis on the existing body of literature that investigates the disambiguation effect.

Methods

Search strategy. We conducted a forward search based on citations of Markman and Wachtel (1988) in Google Scholar, and by using the keyword combination "mutual exclusivity" in Google Scholar (September 2013; November 2017). [1] Additional papers were identified through citations and by consulting experts in the field. We then narrowed our sample to the subset of studies that used one of two different paradigms: (a) an experimenter says a novel word in the context of a familiar object and a novel object and the child guesses the intended referent (the canonical paradigm; "Familiar-Novel"), or (b) experimenter first provides the child with an unambiguous mapping of a novel label to a novel object, and then introduces a second novel object and asks the child to identify the referent of a second novel label ("Novel-Novel"). For Familiar-Novel conditions, we included conditions that included more than one familiar object (e.g. Familiar-Familiar-Novel). From these conditions, we restricted our sample to only those that satisfied the following criteria: (a) participants were children (less than 12 years of age)[2], (b) referents were objects or pictures (not facts or object parts), and (c) no incongruent cues (e.g. eye gaze at familiar object). All papers used either forced-choice pointing or eye-tracking methodology. All papers were peer-reviewed with the exception of two dissertations (Williams, 2009; Frank, I., 1999), but all main results reported

[1] Data and analysis code for this and subsquent studies are available in an online repository at: https://github.com/langcog/me_vocab

[2] This cutoff was arbitrary but allowed us to include conditions from older children from non-typically-developing populations.

below remain the same when these papers are excluded. In total, we identified 43 papers that satisfied our selection criteria and had sufficient information to calculate an effect size.

Coding. For each paper, we coded separately each relevant condition with each age group entered as a separate condition. For each condition, we coded the paper metadata (citation) as well as several potential moderator variables: mean age of infants, method (pointing or eyetracking), participant population type, estimates of mean vocabulary size of the sample population from the Words and Gestures form of the MacArthur-Bates Communicative Development Inventory when available (Fenson et al., 2007, MCDI; 1994), referent type (object or picture), and number of alternatives in the forced choice task. We used production vocabulary as our estimate of vocabulary size since it was available for more studies in our sample. We coded participant population as one of three subpopulations that have studied in the literature: (a) typically-developing monolingual chilldren, (b) multilingual children (including both bilingual and trilingual children), and (c) non-typically developing children. Non-typically developing conditions included children with selective language impairment, language delays, hearing impairment, autism spectrum disorder, and down-syndrome.

In order to estimate effect size for each conditions, we also coded sample size, proportion novel-object selections, baseline (e.g., .5 in a 2-AFC paradigm), and standard deviations for novel object selections, *t*-statistic, and Cohen's *d*. For several conditions, there was insufficient data reported in the main text to calculate an effect size (no means and standard deviations, *t*-statistics, or Cohen's *d*s), but we were able to estimate the means and standard deviations though measurement of plots ($N = 13$), imputation from other data within the paper ($N = 4$; see SI for details), or through contacting authors ($N = 26$). Our final sample included 157 effect sizes ($N_{\text{typical-developing}} = 135$; $N_{\text{multilingual}} = 12$; $N_{\text{non-typically-developing}} = 10$).

Statistical approach. We calculated effect sizes (Cohen's d) from reported means and standard deviations where available, otherwise we relied on reported test-statistics (t or d). Effect sizes were computed by a script, `compute_es.R`, available in the Github repository. All analyses were conducted with the metafor package (Viechtbauer & others, 2010) using mixed-effect models with grouping by paper.[3] In models with moderators, moderators variables were included as additive fixed effects. All estimate ranges are 95% confidence intervals.

Analyses

We conducted a separate meta-analysis for four theoretically-relevant conditions: Familiar-Novel trials with typically developing participants, Novel-Novel trials with typically developing participants, conditions with multilingual participants, and conditions with non-typically developing participants.

Typically-Developing Population: Novel-Familiar Trials. We first examined effect sizes for the disambiguation effect for typically-developing children in the canonical familiar-novel paradigm. This is the central data point that theories of disambiguation must explain.

Results.

The overall effect size for these conditions was 1.1 [0.79, 1.42], and reliably greater than zero (p <.001; Figure 1). The effect sizes contained considerable heterogeneity, however (Q = 968.13; p <.001).

We next tried to predict this heterogeneity with two moderators corresponding to developmental change: age and vocabulary size. In a model with age as a moderator, age

[3] The exact model specification was as follows: `metafor::rma.mv(yi = effect_size, V = effect_size_var, random = ~ 1 | paper)`.

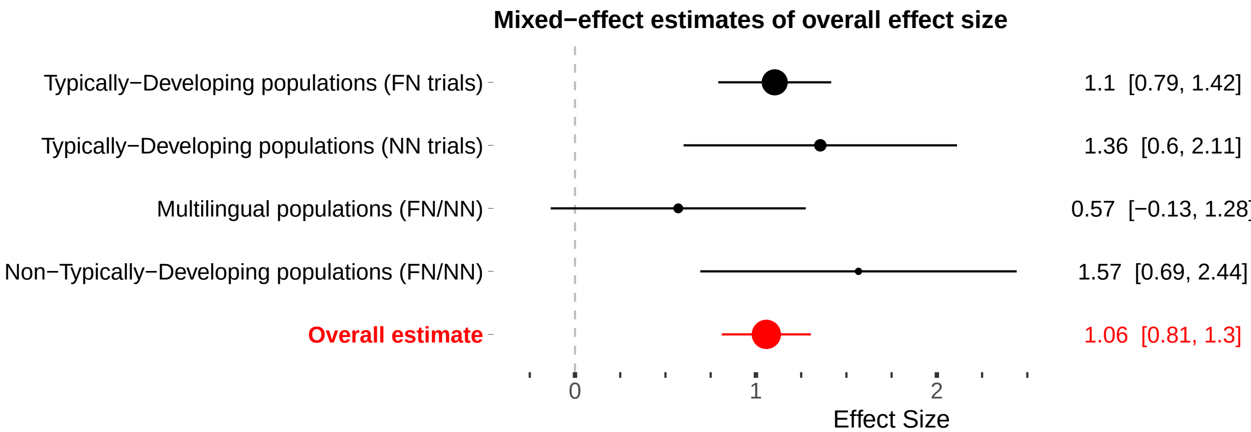

Figure 1. Mixed-effect effect size estimates for all conditions (red) and each of the four theoretically-relevant conditions in our sample. Ranges are 95 percent confidence intervals. Point size corresponds to sample size. FN = Familiar-Novel trials; NN = Novel-Novel trials.

was a reliable predictor of effect size (β = 0.05, z = 11.85, p <.001; see Table 1), suggesting that the disambiguation effect becomes larger as children get older. Age of participants was highly correlated with vocabulary size in our sample (r = 0.65, $p < .01$), so next we asked whether vocabulary size predicted independent variance in the magnitude of the disambiguation bias on the subset of conditions for which we had estimates of vocabulary size (N = 23). To test this, we fit a model with both age and vocabulary size as moderators. Vocabulary size (β = 0.07, z = 2.14, p = 0.03), but not age (β = -0.78, z = -1.11, p = 0.27, was a reliable predictor of disambiguation effect size.

These analyses confirm that the disambiguation phenomenon is robust, and associated with a relatively large effect size (d = 1.1 [0.79, 1.42]). In addition, this set of analyses provides theory-constraining evidence about the mechanisms underlying the effect. In particular, the finding that vocabulary predicts more variance in effect size, compared to age, suggests that there is an experience related component to the mechanism, independent of maturational development alone.

Table 1

Meta-analytic model parameters for model including age as a fixed effect. The first model (top) estimates effect sizes for all studies in our sample. The four subsequent models present separate models paremeters for four seperate conditions. Ranges are 95 percent confidence intervals.

Model	n	term	estimate	Z	p
Overall estimate	157	intercept	-0.18 [-0.47, 0.11]	-1.21	0.23
		age	0.03 [0.03, 0.04]	11.32	<.01
Typically-Developing populations (FN trials)	117	intercept	-0.33 [-0.71, 0.05]	-1.73	0.08
		age	0.05 [0.04, 0.05]	11.85	<.01
Typically-Developing populations (NN trials)	18	intercept	0.06 [-0.8, 0.93]	0.15	0.88
		age	0.03 [0.01, 0.04]	3.55	<.01
Multilingual populations (FN/NN)	12	intercept	0.05 [-0.78, 0.87]	0.11	0.91
		age	0.02 [0, 0.03]	1.77	0.08
Non-Typically-Developing populations (FN/NN)	10	intercept	-0.58 [-2.08, 0.92]	-0.75	0.45
		age	0.04 [0.01, 0.06]	3.15	<.01

Note. n = sample size (number of studies); FN = Familiar-Novel; NN = Novel-Novel.

Typically-Developing Population: Novel-Novel Trials. The results from the Familiar-Novel trials point to a role for vocabulary knowledge in the strength of the disambiguation effect. One way in which this vocabulary knowledge could lead to increased performance on the Familiar-Novel disambiguation task is through increased certainty about the label associated with the familiar word: If a child is less certain that a ball is called "ball," then the child should be less certain that the novel label applies to the novel object. Novel-Novel trials control for potential variability in certainty about the familiar object by teaching participants a new label for a novel object prior to the critical disambiguation trial, where this previously-learned label becomes the "familiar" object in the disambiguation trial.

If knowledge of the familiar object is not the only contributor to age-related changes in the disambiguation effect, then there should be developmental change in Novel-Novel trials, as well as Novel-Familiar trials. In addition, if the strength of knowledge of the "familiar" object influences the strength of the disambiguation effect, then the overall effect size should be smaller for Novel-Novel trials, compared to Familiar-Novel trials.

For conditions with the Novel-Novel trial design, the overall effect size was 1.36 [0.6, 2.11] and reliably greater than zero (p <.001). We next asked whether age predicted some of the variance in these trials by fitting a model with age as a moderator. Age was a reliable predictor of effect size (β = 0.03, z = 3.55, p <.001), suggesting that the strength of the disambiguation bias increases with age. There were no Novel-Novel conditions in our dataset where the mean vocabulary size of the sample was reported, and thus we were not able to examine the moderating role of vocabulary size on these trials.

Finally, we fit a model with both age and trial type (Familiar-Novel or Novel-Novel) as moderators of the disambiguation effect. Both moderators predicted independent variance in disambiguation effect size (age: β = -0.08, z = -0.42, p = 0.68; trial-type: β = 0.04, z = 12.34, p <.0001), with Familiar-Novel conditions and conditions with older participants tending to have larger effect sizes.

These analyses point to an influence on the disambiguation effect of both development (either via maturation or experience-related changes) as well as the strength of the familiar word representation. A successful theory of disambiguation will need to account for both of these empirical facts.

Multilingual Population. We next turn to a different population of participants: Children who are simultaneously learning multiple languages. This population is of theoretical interest because it allows us to isolate the influence of linguistic knowledge from the influence of domain-general capabilities. If the disambiguation phenomenon relies on mechanisms that are domain-general and independent of linguistic knowledge, then we should expect the magnitude of the effect size to be the same for multilingual children compared to monolingual children.

Children learning multiple languages reliably showed the disambiguation effect (d = 1.57 [0.69, 2.44]). We next fit a model with both monolingual (typically-developing) and multilingual participants, predicting effect size with language status (monolingual vs. multilingual), while controlling for age. Language status was not a reliable predictor of effect size ($\beta = 0.20$, $z = 1.42$, $p = 0.16$), but age was ($\beta = 0.03$, $z = 11.54$, $p < .0001$).

In sum, these data do not provide strong evidence that language-specific knowledge influences effect size. However, the small sample size of studies from this population limit the power of this model to detect a difference if one existed.

Non-Typically-Developing Population. Finally, we examine a third-population of participants: non-typically developing children. This group includes a heterogenous sample of children with diagnoses including Autism-Spectrum Disorder (ASD), Mental Retardation, Williams Syndrome, Late-Talker, Selective Language Impairment, and deaf/hard-of-hearing. These populations are of theoretical interest because they allow us to observe how impairment to a particular aspect of cognition influences the magnitude of the disambiguation effect. For example, children with ASD are thought to have impaired social reasoning skills (e.g., Phillips, Baron-Cohen, & Rutter, 1998); thus, if children with ASD are able to succeed on disambiguation tasks, to a first approximation this information might suggest that social reasoning skills are not critically involved in making disambiguation inferences (de Marchena, Eigsti, Worek, Ono, & Snedeker, 2011; Preissler & Carey, 2005).

Overall, non-typically developing children succeeded on disambiguation tasks (d = 1.57 [0.69, 2.44]). In a model with age as a moderator, age was a reliable predictor of the effect, suggesting children became more accurate with age, as with other populations ($\beta = 0.04$, z

= 3.15, p <.001). We were not able to examine the potential moderating role of vocabulary size for this population because there were only 3 conditions where mean vocabulary size was reported.

We also asked whether the effect size for non-typically developing children differed from typically-developing children, controlling for age. We fit a model predicting effect size with both development type (typical vs. non-typical) and age. Development type was a reliable predictor of effect size with non-typically developing children tending to have a smaller bias compared to typically developing children (β = -0.50, z = -2.86, p <.0001). Age was also a reliable predictor of effect size in this model (β = 0.04, z = 11.34, p <.0001).

This analysis suggests that non-typically developing children succeed in the disambiguation paradigm just as typically developing children do, albeit at lower rates. Theoretical accounts of the disambiguation phenomenon will need to account for how non-typically developing children are able to succeed in the disambiguation task, despite a range of different cognitive impairments.

Discussion

To summarize our meta-analytic findings, we find a robust disambiguation effect in each of the three populations we examined, as well as evidence that the magnitude of this effect increases across development. We also find that the effect is larger in the canonical Novel-Familiar paradigm compared to the Novel-Novel paradigm, but both designs show roughly the same developmental trajectory.

Taken together, these analyses provide several theoretical constraints with respect to the mechanism underlying the disambiguation effect. First, language experience likely accounts for some developmental change. This conclusion derives from the fact that we see a larger effect size in Novel-Familiar trials compared to Novel-Novel trials, and that there is a suggestive correlation between vocabulary size and the strength of the disambiguation effect. Second, independent of familiar word knowledge, the strength of the bias increases across development. This constraint comes from the fact that the bias strengthens across development in the Novel-Novel conditions, and from the fact that there is not a significant impairment to effect in multilingual children (who presumably have less language experience with any particular language). Third, children with a range of different impairments are able to make the inference, suggesting that no single mechanism is both necessary and sufficient for the effect.

In the next section, we gather additional evidence to shed light on the relative contributions of these different mechanisms on the disambiguation effect. In particular, we use experimental methods to more directly examine the relationship between linguistic experience and the disambiguation effect.

Experiment 1: Disambiguation Effect and Vocabulary Size

Our meta-analysis points to a robust developmental increase in the strength of the disambiguation effect with age. However, the meta-analytic approach is limited in its ability to measure this relationship since few studies in our sample measure vocabulary size (N = 8), and even fewer measure vocabulary size at multiple ages within the same study (Markman et al., 2003; $N = 2$; Mather & Plunkett, 2009). In Experiment 1, we therefore aimed to test the prediction that children with larger vocabularies should have a stronger disambiguation bias by measuring vocabulary size in a large sample of children across multiple ages who also completed the disambiguation task. We find that vocabulary size is a strong predictor of the strength of the disambiguation effect across development and that vocabulary size predicts more variance than developmental age.

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants. To determine sample size, we conducted a power analyses and determined that we needed a large N to do partial correlations with age. and so that meant we needed a short and easy testing procedure, motivating us to do everything on tablet in the museum… then after some piloting we selected words that we estimated (on the basis of Wordbank and other resources) would be challenging for children across the broader age range. Of course, this assessment is way worse than CDI but we couldn't do that in the full developmental range of the study, and a full PPVT would be prohibitively time consuming.

A sample of 226 children were recruited at the Children's Discovery Museum of San Jose. 72 children were excluded because they did not satisfy our planned inclusion criteria: within the age range of 24-48 months (n excluded = 13), completed all trials (n excluded= 48), exposed to English greater than 75% of the time (n excluded = 37), and correctly answered at least half of the familiar noun control trials (n excluded = 55). Our final sample included 154 children ($N_{\text{females}} = 93$).

Stimuli. The disambiguation task included color pictures of 14 novel objects (e.g., a funnel) and 24 familiar objects (e.g. a ball; see Appendix). The novel words were the real 1-2 syllables labels for the unfamiliar objects (e.g., "funnel", "tongs", etc.; See Appendix). Items in the vocabulary assessment were a fixed set of 20 developmentally appropriate words from the Pearson Peabody Vocabulary Test (see Appendix; L. M. Dunn, Dunn, Bulheller, & Häcker, 1965).

Design and Procedure. Sessions took place individually in a small testing room away from the museum floor. The experimenter first introduced the child to "Mr. Fox," a cartoon character who wanted to play a guessing game (see Fig. 1). The experimenter explained that Mr. Fox would tell them the name of the object they had to find, so they had to listen carefully. Children then completed a series of 19 trials on an iPad, 3 practice trials followed by 16 experimental trials. In the practice trials, children were shown two familiar pictures (FF) on the tablet and asked to select one, given a label (e.g. "Touch the ball!"). If the participant chose incorrectly on a practice trial, the audio would correct them and allow the participant to choose again. The audio was presented through the tablet speakers.

The child then completed the test phase. Each test trial consisted of two screens: One presenting a single object and an unambigious label (Fig. Xb), and another presenting two objects and a single label (Fig. Xc). The child's task was to identify the referent on the second screen. Within participants, we manipulated two features of the task: the target referent (Novel (Experimental) or Familiar (Control)) and the type of alternatives (Novel-Familiar or Novel-Novel; NF or NN). On novel referent trials (Experimental), children were expected to select a novel object via the disambiguation inference. On familiar referent trials (Control), children were expected to select the correct familiar object. On Novel-Familiar trials, children saw a picture of a novel object and a familiar object (e.g. a funnel and a ball). On Novel-Novel trials, children saw pictures of two novel objects (e.g. a pair of tongs and a leak). The design features were fully crossed such that half of the trials were of each trial type (Experimental-NF, Experimental-NN, Control-NF, Control-NN; Table 2). Trials were presented randomly, and children were only allowed to make one selection.

After the disambiguation task, we measured children's vocabulary in a simple vocabulary assessment. in which children were presented with four randomly selected images and prompted to choose a picture given a label. Children completed two practice trials followed by 20 test trials.

Data analysis. Selections on the disambiguation task were coded as correct if the participant selected the familiar object on Control and the novel object on Experimental

Table 2

Design for each of the four trial types. "N" indicates a novel referent and "F" indicates a familiar referent. Each test trial involed two displays. The first introduced an object and its label unabigiously. The second presented two objects and a single label and children were asked to identify the target referent.

Trial Type	Screen 1 Display	Screen 2 Display	Target (Audio)
Experimental	F	NF	N
Experimental	N_1	N_1N_2	N_2
Control	F	NF	F
Control	N_1	N_1N_2	F

trials. We centered both age and vocabulary size for interpretability of coefficients. All models are logistic mixed effect models fit with the lme4 package in R (D. Bates, Mächler, Bolker, & Walker, 2015). Each model was fit with the maximal random effect structure. All ranges are 95% confidence intervals. Effect sizes are Cohen's *d* values.

Results and Discussion

Participants completed the three practice trials (FF) with high accuracy, suggesting that they understood the task (M = 0.91 [0.87, 0.94]).

We next examined performance on the four trial types. Children were above chance (.5) in both types of control conditions where they were asked to identify a familiar referent (Control-NF: M = 0.89, SD = 0.17, d = 2.35 [2.06, 2.64]; Control-NN: M = 0.78, SD = 0.25, d = 1.14 [0.9, 1.38]). Critically, children also succeeded on both types of experimental trials where they were required to select the novel object (NF: M = 0.84, SD = 0.21, d = 1.61 [1.35, 1.87]; NN: M = 0.77, SD = 0.28, d = 0.95 [0.71, 1.19]).

To compare all four conditions, we fit a model predicting accuracy with target type (F (Control) vs. N (Experimental)) and trial type (NF vs. NN) as fixed effects. We included both target type and trial type as main effects as well as a term for their interaction. There was a main effect of trial type, suggesting that participants were less accurate in NN trials compared to NF trials (B = -0.87, SE = 0.25, Z = -3.51, $p < .001$). The main effect of target type was not significant (B = -0.49, SE = 0.29, Z = -1.69, p = 0.09). The interaction between the two factors was marginal (B = 0.57, SE = 0.36, Z = 1.56, p = 0.12), suggesting that Novel target trials (Experimental) were more difficult than Familiar target trials (Control) for NF trials but not NN trials.

Our main question was how accuracy on the experimental trials changed over development. We examined two measures of developmental change: Age (months) and vocabulary size, as measured in our vocabulary assessment. We assigned a vocabulary score to each child as the proportion correct selections on the vocabulary assessment out of 20 possible. Age and vocabulary size were positively correlated, with older children tending to have larger vocabularies compared to younger children ($r = 0.43$ [0.29, 0.55], $p < .001$).

Figure 3 shows log linear model fits for accuracy as a function of age (left) and vocabulary size (right) for both NF and NN trial types. To examine the relative influence of maturation and vocabulary size on accuracy, we fit a model predicting accuracy with vocabulary size, age, and trial type (Experimental-NN, and Experimental-NF). We included all possible main and interaction terms as fixed effects. Table 2 presents the model parameters. The only reliable predictor of accuracy was vocabulary size (B = 5.93, SE = 0, Z = 6406.33, $p <.0001$), suggesting that children with larger vocabularies tended to be more accurate in the disambiguation task. Notably, age was not a reliable predictor of accuracy over and above vocabulary size (B = 0.02, SE = 0, Z = 21.8, $p <.0001$).

Discussion. Experiment 1 directly examines the relationship between the strength of the disambiguation effect and vocabulary size. We find that the strength of the

Table 3

Parameters of logistic mixed model predicting accuracy on disambiguation trials as a function of trial type (Novel-Familiar (NF) vs. Novel-Novel (NN), age (months), and vocabulary size as measured by our vocabulary assessment.

term	Beta	SE	Z	p
(Intercept)	2.01	0.00	2,240.62	$<.0001$
Vocabulary	5.93	0.00	6,406.33	$<.0001$
Trial Type (NN)	-0.51	0.00	-564.56	$<.0001$
Age	0.02	0.00	21.80	$<.0001$
Vocabulary x Trial Type (NN)	-2.95	0.00	-3,185.91	$<.0001$
Vocabulary x Age	-0.01	0.00	-9.88	$<.0001$
Age x Trial Type (NN)	0.02	0.00	18.24	$<.0001$
Vocabulary x Age x Trial Type (NN)	0.13	0.00	145.54	$<.0001$

disambiguation effect is highly predicted by vocabulary size. In addition, we find that the bias is larger for NF trials, compared to NN trials.

The pattern of findings is broadly consistent with meta-analytic estimates of those same effects. Figure 4 presents the data from the experimental conditions in Experiment 1 together with meta-analytic estimates, as a function of age. To compare the experimental data with the meta-analytic data, an effect size was calculated for each participant.[4] As in the meta-analytic models, the effect size is smaller for NN trials compared to NF trials, though the magnitude of this difference is smaller. We also see that the variance is larger for the meta-analytic estimates compared to the experimental data, presumably because there is

[4]Because some participants had no variability in their responses (all correct or all incorrect), we used the across-participant mean standard deviation as an estimate of the participant level standard deviation in order to convert accuracy scores into Cohen's d values.

more heterogeneity across experiments than across participants within the same experiment. The experimental data thus provide converging data with the meta-analysis that there is developmental change in the strength of the bias, and that the effect is weaker for NN trials.

However, there are some notable differences between the meta-analytic results and that of Experiment 1. the developmental effect is smaller, etc. The directions and significances are the same but the magnitude suggests that there are other things perhaps going on, e.g. 1) people adjusting paradigms for different age groups, 2) different populations, 3) different familiar words, etc. etc. I would add a paragraph discussing these differences.

In addition, the data from Experiment 1 provide new evidence relevant to the mechanism underlying the effect: children with larger vocabulary tend to to have a stronger disambiguation bias. In principle there are two ways that vocabulary knowledge could support the disambiguation inference. The first is by influencing the strength of the learner's knowledge about the label for the familiar word: If a learner is more certain about the label for the familiar object, they can be more certain about the label for novel object. This account explains the developmental change observed for NF trials. However, this account does not explain the relationship of vocabulary with NN trials, since no prior vocabulary knowledge is directly relevant to this inference. This relationship between vocabulary size and NF size suggests that vocabulary knowledge could also influence the effect by providing evidence for general constraint that there is a one-to-one mapping between words and referents. This empirical fact is consistent with the overhypothesis account.

Importantly, however, data from both the meta-analytic study and the current experiment only provide correlational evidence about the relationship between vocabulary size and the disambiguation inference. In Experiment 2, we experimentally test the hypothesis that the strength of the learner's knowledge about the familiar object influences the strength of the disambiguation inference, thereby testing one possible route through which vocabulary knowledge may be related to the disambiguation phenomenon.

Table 4

Demographics of children in Experiment 2.

Age group	Mean age (months)	Sample size
2	30.99	38
3	40.99	35
4	52.16	37

Experiment 2: Disambiguation Effect and Familiarity

In Experiment 2, we test a causal relationship between vocabulary size and the disambiguation effect by experimentally manipulating the strength of word knowledge. We do this by teaching participants a label for a novel object and varying the number of times the object is labeled. This manipulation allows us to vary children's certainty about the label for an object, with objects that have been labeled more frequently associated with high certainty about the label name. The newly, unabiguously labeled object then serves as the "familiar" object in a novel-novel trial. If the strength of vocabulary knowledge about the "familiar" object influences, the strength of the disambiguation effect, then we should expect a larger bias when the the familiar object has been labeled more frequently. We find a pattern consistent with this prediction.

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants. We planned a total sample of 108 children, 12 per between-subjects labeling condition, and 36 total in each one-year age group (see Table 3). Our final sample was 110 children, ages 25 – 58.50 months, recruited from the floor of the Boston Children's

Museum. Children were randomly assigned to the one-label, two-label, or three label condition, with the total number of children in each age group and condition ranging between 10 and 13.

Materials. Materials were the set of novel objects used in de Marchena et al. (2011), consisting of unusual household items (e.g., a yellow plastic drain catcher) or other small, lab-constructed stimuli (e.g., a plastic lid glued to a popsicle stick). Items were distinct in color and shape.

Procedure. Each child completed four trials. Each trial consisted of a training and a test phase in a "novel-novel" disambiguation task (de Marchena et al., 2011). In the training phase, the experimenter presented the child with a novel object, and explicitly labeled the object with a novel label 1, 2, or 3 times ("Look at the *dax*"), and contrasted it with a second novel object ("And this one is cool too") to ensure equal familiarity. In the test phase, the child was asked to point to the object referred to by a second novel label ("Can you show me the *zot*?"). Number of labels used in the training phase was manipulated between subjects. There were eight different novel words and objects. Object presentation side, object, and word were counterbalanced across children.

Data analysis. We followed the same analytic approach as we registered in Experiment 1, though data were collected chronologically earlier for Experiment 2. Responses were coded as correct if participants selected the novel object at test. A small number of trials were coded as having parent or sibling interference (N = 11), experimenter error (N = 2), or a child who recognized the target object (N = 4), chose both objects (N = 2) or did not make a choice (N = 8). These trials were excluded from further analyses; all trials were removed for two children for whom there was parent or sibling interference on every trial. We centered both age and number of labels for interpretability of coefficients. The analysis we report here is consistent with that used in Lewis and Frank (2013), though there are some slight numerical differences due to reclassification of exclusions.

Table 5

Parameters of logistic mixed model predicting accuracy on disambiguation trials as a function of age (months) and number of times a label for the familiar object was observed.

term	B	SE	Z	p
(Intercept)	0.31	0.10	2.94	$< .001$
Age	0.05	0.01	4.13	$< .001$
Num. Labels Observed	0.48	0.13	3.75	$< .001$
Age x Num. Labels Observed	0.02	0.01	1.58	0.11

Results and Discussion

As predicted, children showed a stronger disambiguation effect as the number of training labels increased, and as noise decreased with age (Figure 5).

We analyzed the results using a logistic mixed model to predict correct responses with age, number of labels, and their interaction as fixed effects, and participant as a random effect. Model results are shown in Table 4. There was a significant effect of age such that older children showed a stronger disambiguation bias and a significant effect of number of labels, such that more training labels led to stronger disambiguation, but the interaction between age and number of labels was not significant.

[Add a caveat in the results and discussion that we can't directly compare magnitudes between the experience effect that we observe in this experiment and the effects of vocabulary size! at best we show that the causal arrow points in the same direction as the correlational one. This is in some sense cold comfort, but at least it's something. :)]

These data provide causal evidence that the strength of knowledge of the familiar word influences the strength of the disambiguation effect. It thus points to one route through which a child's vocabulary knowledge might influence the disambiguation inference.

General Discussion

These three constraints are consistent with many of individual proposed accounts, as well as a various combinations of them. For example, an effect of language experience on the disambiguation effect via vocabulary knowledge is most consistent with the overhypothesis account, which predicts a stronger learned bias with vocabulary development. However, all four accounts predict developmental change in the NN trials. Under the overhypothesis account, as children are exposed to more language, they develop a stronger learned bias even when the "familiar" word is not previously known; Under the pragmatics account, as children are exposed to more language, they develop more skill in making social inferences, which would led to increased performance on the NN trials; And, under the bias and probabilistic accounts, maturational change could contribute to development in domain-general abilities, leading to a stronger disambiguation inference. Finally, the ability of children to succeed in the disambiguation tasks despite a range of impairments suggests that accounts that rely on a single mechanism, such as pragmatic reasoning or a mutual exclusivity constraint alone, are unlikely to describe the mechanism underlying the disambiguation effect across all children.

We conclude that:

1. Explanations of ME are not themselves mutually exclusive and likely more than one is at play.
2. The balance of responsibility for behavior likely changes developmentally, with basic biases playing a greater role for younger children and learned overhypotheses playing a greater role for older children.
3. All existing accounts put too little emphasis on the role of experience and strength of

representation; this lack of explicit theory in many cases precludes definitive tests of mechanism.

To summarize, the empirical findings that a successful theory of ME must account for are:

(1) Why the effect is present in young children, but gets larger with development (developmental change);
(2) How language experience supports the effect (multilingualism evidence);
(3) Why pragmatic reasoning can support the effect (speaker change evidence), but why it is not necessary (autism evidence).

The role of development in theories of the ME effect

What are the cognitive mechanisms underlying the ME effect? A number of proposals have been made in the literature, many of which overlap or differ only in subtle ways. Here we briefly describe several influential proposals, highlighting the commonalities and differences across theoretical views. This review is necessarily selective; in addition in this portion of the manuscript, we engage primarily with qualitative and verbal theories. We return to the issue of computational models that treat the ME effect in a later section.

Constraint and bias accounts. One proposal is that children have a constraint or bias that is innate or emerges after very limited language input. Under one version of this account (Markman & Wachtel, 1988; Markman et al., 2003), children have a constraint on the types of lexicons considered when learning the meaning of a new word – a "mutual exclusivity constraint." Under this constraint, children are biased to consider only those lexicons that have a one-to-one mapping between words and objects. Importantly, this constraint is probabilistic and thus can be overcome in cases where it is incorrect (e.g., property names or super-/sub-ordinate labels), but it nonetheless serves to restrict the set of lexicons initially entertained when learning the meaning of a novel word. In principle, this constraint could be the result of either domain-specific or domain-general processes (Markman, 1992). As a domain general property, the ME constraint could be related to other cognitive mechanisms that lead learners to prefer one-to-one mappings (e.g. blocking and overshadowing in classical condition and the discounting principle in motivational research; Lepper, Greene, and Nisbett (1973)).

As formulated by Markman and colleagues, the ME constraint operates at the level of extensions (objects), not concepts. For example, the ME constraint says that the labels "policeman" and "cop" – referring to the same entity in the world – are violations of the constraint. Similarly, terms at different levels of the semantic hierarchy that can have the same extensions, such as "animal" and "dog," are also seen as ME violations. In contrast, these cases are not violations in theories that posit the explanatory construct at the level of concepts (e.g., pragmatic accounts). The distinction between concepts and objects in each theoretical view is important for evaluating whether empirical evidence is consistent with a proposal. Note, however, that in the canonical ME paradigm, where the two referents are both different concepts and objects, the accounts at both levels make identical predictions.

A related proposal to the ME constraint is that children have a bias to map novelty to novelty (Novel-Name Nameless-Category principle (N3C); Golinkoff et al., 1994; Mervis & Bertrand, 1994). This principle differs from the ME constraint in that the rejection of the familiar object as a potential referent is not part of the inference; instead, children are argued only to map the two novel elements to each other, the novel label and the object (thereby only implicitly rejecting the the familiar object as a referent for the novel label). The N3C principle is argued to be domain-specific to language.

Under a third account, children are motivated to identify objects for which they do not know a label for and fill the "lexical gap" with the novel label (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Merriman, Bowman, & MacWhinney, 1989).

Probabilistic accounts. Probabilistic accounts contend that the ME effect does not derive from an explicit representation related to the one-to-one regularity, as proposed by the constraints and bias accounts. Rather, under these accounts, the effect is the product of a word learning system that tracks the frequency of exemplars of words and their referents over time, and then reasons probabilistically about the most likely referent for a novel word within the referential context (Fazly, Alishahi, & Stevenson, 2010; M. C. Frank, Goodman, & Tenenbaum, 2009; Kachergis, Yu, & Shiffrin, 2012; McMurray, Horst, & Samuelson, 2012; Regier, 2005). These accounts are typically posed in the context of more explicit computational models and we defer discussion of these until later in the manuscript

Pragmatic accounts. Under pragmatic accounts, the ME effect derives from reasoning about the intention of the speaker within the referential context (Clark, 1987, 1988, 1990; Diesendruck & Markson, 2001). The critical aspect of this account is the claim that children assume that "every two forms contrast in meaning" (Clark, 1988, p. 417), or the "Principle of Contrast." Clark also argues that speakers hold a second assumption – that speakers within the same speech community use the same words to refer to the same objects ("Principle of Conventionality"). The ME effect then emerges from the interaction of these two principles. That is, the child reason's implicitly: You used a word I've never heard before. Since, presumably we both call a ball "ball" and if you'd meant the ball you would have said "ball," this new word must refer to the new object. Clark (1988, 1990) argues that these two principles are learned, but emerge from a more general understanding that other people have intentions (Grice, 1975; Tomasello, Carpenter, Call, Behne, & Moll, 2005).

Logical inference accounts. Halberda (2003) argues that the ME effect is the result of domain-general processes used for logical reasoning. Under this proposal, children are argued to be solving a disjunctive syllogism ("A or B, not A, therefore B") by rejecting labels for known objects. For example, upon hearing the novel label "dax," the child would implicitly reason that the referent could be either object A or B, and then reject object A because it already has a known label. By deduction, the child would then conclude that "dax" refers to object B. This account shares the same formal reasoning structure as pragmatic accounts, but differs in the underlying source of the key inference: While pragmatic accounts argue that children conclude that object B must be the referent on the basis of reasoning about intention, the logical inference account proposes that this same inference is made on the basis of logical reasoning.

Over-hypothesis accounts. Lewis and Frank (2013) suggest that the ME effect could emerge by learning from the statistics of the child's linguistic. That is, given evidence that words tend to refer to a single concept, the child might develop a learned "overhypothesis" (Kemp, Perfors, & Tenenbaum, 2007) that the lexicon is structured such that each concept is associated with one and only one label. The learning mechanisms are argued to be probabilistic and domain general, while the learned overhypothesis is specific to the structure of the lexicon. The emergent overhypothesis about the structure of the lexicon would be similar to the knowledge a learner is proposed to have under the constraints and biases account.

In order for learning to get off the ground, however, children must notice the one-to-one mapping between a word and a concept in the context of a particular instances of a label's usage. Lewis and Frank (2013) suggest that this ability could derive from a variety of different mechanisms that make use of the structure of the learning task, such as pragmatic, probabilistic, and logical inference accounts. Merriman (1986) make a similar proposal, but argue that the overhypothesis is learned primarily from explicit parental corrections (e.g., "that's an apple, not an orange").

Under the overhypothesis account, then, the ME effect emerges from multiple mechanisms at two different timescales – one as a function of information about the pragmatic or inferential structure of the communicative context and one as a function of learned higher-order knowledge about how the lexicon is structured. Both mechanisms would then contribute to the inference with different weights across development and across children.

An important starting point for our analysis is to observe that theoretical accounts of mechanisms underlying the ME effect that have proposed in the literature are not mutually exclusive with each other (Momen & Merriman, 2002). As argued by (**???**), testing different mechanisms in isolation is the result of an experimental approach to theory building, rather than a reflection of an assumption that there exists one and only one mechanism underlying the effect. That is, in order to identify whether a mechanism is *sufficient* to give rise to the ME effect, logical researchers design experiments in which the ME effect can be observed only if a particular cognitive mechanism is sufficient for the effect. If the effect is observed under these conditions, it provides evidence only that the mechanism is sufficient for the effect, but not that it is necessary and not that other mechanisms are not also sufficient. Indeed, there is reason to think that redundancy in mechanisms for the same behavior is a desirable property of a cognitive system.

Instead, we argue that multiple mechanisms likely support the ME effect. Each child may be making use of multiple mechanisms with varying weights across development and situations, and the relative weights of these different mechanisms may vary across children. For example, learners may be making use of both general knowledge about how the lexicon is structured as well as information about the pragmatic or inferential structure of the task, and both of these sources of information support the ME inference.

Potential sources of developmental change Cognitive limitations mean that you just don't think of the other object when one is named (Merriman 1986b) - better at coordinating concepts (related to Flavell, though not the same. . . .he actually argues the opposite).. Perceptual experience - Merriman 1986b Linguistic input Mervis 1987 - not till children accept parents authority Time constraint - Frank (Halberda)

References

---

nocite: | Markman (1992)

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

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.

Bolinger, D. (1977). Meaning and form.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, *6*(3), 184–205.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of Language Acquisition. Hillsdale, NJ: Erlbaum*.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(2), 317–335.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language*, *17*(2), 417–431.

de Marchena, A., Eigsti, I., Worek, A., Ono, K., & Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, *119*(1), 96–113.

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

Dunn, L. M., Dunn, L. M., Bulheller, S., & Häcker, H. (1965). *Peabody Picture Vocabulary Test*. American Guidance Service Circle Pines, MN.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *MacArthur-Bates Communicative Development Inventories*. Paul H. Brookes Publishing Company.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Golinkoff, R. M., Hirsh-Pasek, K., Baduini, C., & Lavallee, A. (1985). 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., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28*(1), 99.

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

Grice, H. (1975). Logic and conversation. *1975*, 41–58.

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

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

Hutchinson, J. (1986). Children's sensitivity to the contrastive use of object category terms.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin & Review*, *19*(2), 317–324.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307–321.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*(1), 129.

Lewis, M., & Frank, M. C. (2013). Modeling disambiguation in word learning via multiple probabilistic constraints. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.*

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

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

Markman, E. M., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Markman, E. M., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*(6619), 813–815.

Mather, E., & Plunkett, K. (2009). Learning words over time: The role of stimulus repetition in mutual exclusivity. *Infancy*, *14*(1), 60–76.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831.

Merriman, W. E. (1986). Some reasons for the occurrence and eventual correction of children's naming errors. *Child Development*, 942–952.

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.

Mervis, C., Golinkoff, R. M., & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic-level category. *Child Development*, *65*(4), 1163–1177.

Momen, N., & Merriman, W. E. (2002). Two-year-olds' expectation that lexical gaps will be filled. *First Language*, *22*(3), 225–247.

Phillips, W., Baron-Cohen, S., & Rutter, M. (1998). Understanding intention in normal development and in autism. *British Journal of Developmental Psychology*, *16*(3), 337–348.

Preissler, M., & Carey, S. (2005). The role of inferences about referential intent in word learning: Evidence from autism. *Cognition*, *97*(1), B13–B23.

Quine, W. (1960). *Word and object* (Vol. 4). The MIT Press.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819–865.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). In search of the uniquely human. *Behavioral and Brain Sciences*, *28*(5), 721–727.

Viechtbauer, W., & others. (2010). Conducting meta-analyses in r with the metafor package. *J Stat Softw*, *36*(3), 1–48.

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.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

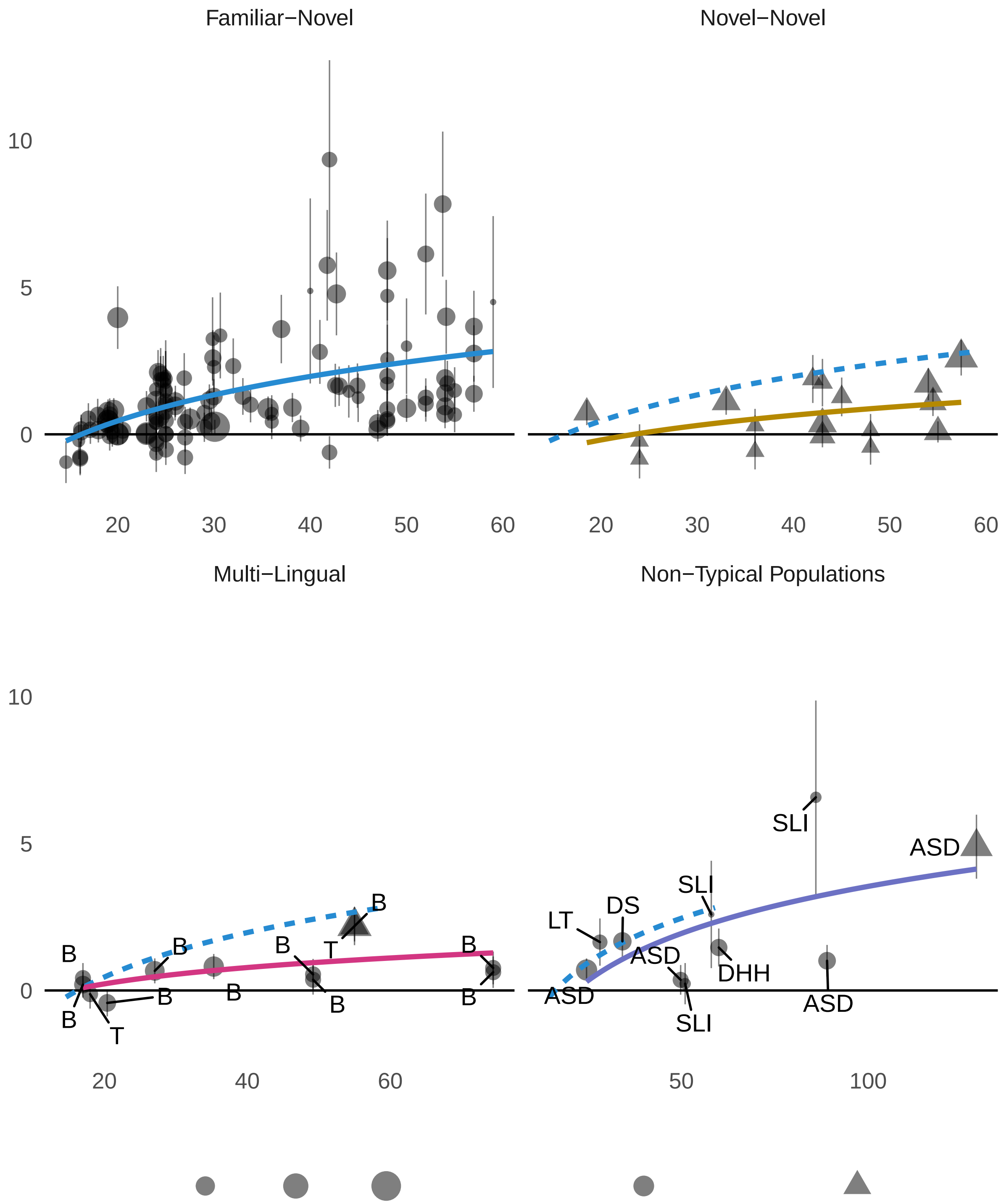

Figure 2. Developmental plots for each moderator. Ranges correspond to 95% confidence intervals. Model fits are log-linear. Point size corresponds to sample size, and point shape corresponds to trial type (Familiar-Novel vs. Novel-Novel). Note that the x-axis scale varies by facet. B = bilingual; T = trilingual; LT = late-talker; ASD = autism spectrum disorder; DS = down syndrome; SLI = selective language imparement; DHH = deaf/heard-of-hearing.

Figure 3. Example screenshots for a Experimental Novel-Familiar test trial. On each test trial, Mr. Fox first appeared to get the child's attention (a). Next, an object appeared and was labeled through the tablet speakers ('It's a ball¡; b). Two objects then appeared and children were asked to make a selection ('Touch the funnel¡; c).

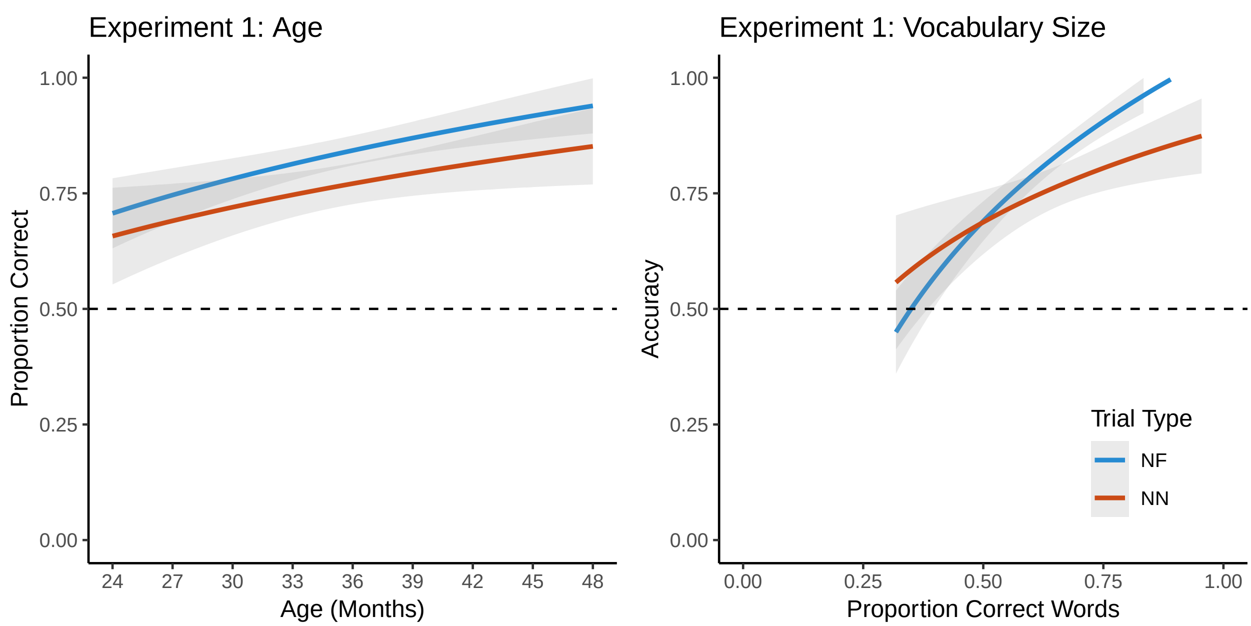

Figure 4. Experiment 1 results. Accuracy as a function of age (months; left) and vocabulary size (proportion correct on vocabulary assessment; right). Blue corresponds to trials with the canonical novel-familiar disambiguation paradigm, and red corresponds to trials with two novel alternatives, where a novel of label for one of the objects is unambiguously introduced on a previous trial. The dashed line corresponds to chance. Ranges are 95% confidence intervals.

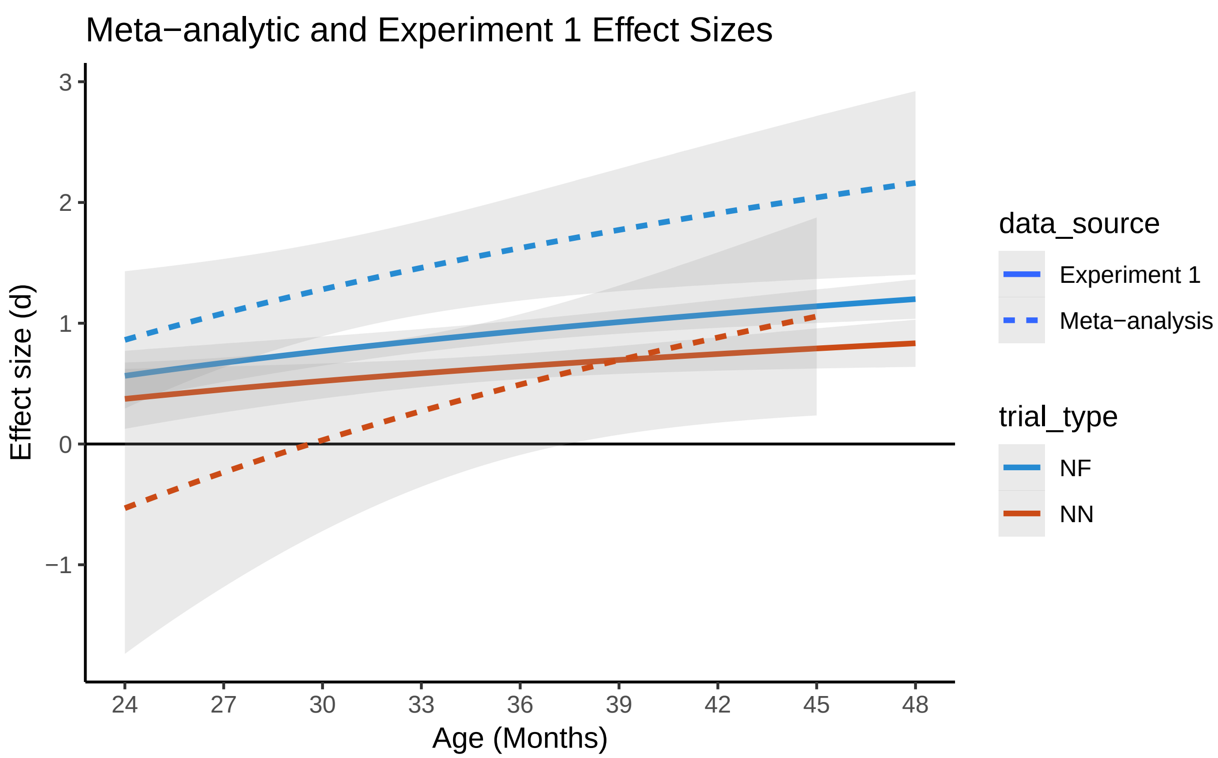

Figure 5. Meta-analytic data and data from experimental trials in Experiment 1 as a function of age. Effect sizes for Experiment 1 data are calculated for each participant, assuming the across-participant mean standard deviation as an estimate of the participant level standard deviation. Ranges are 95% confidence intervals.

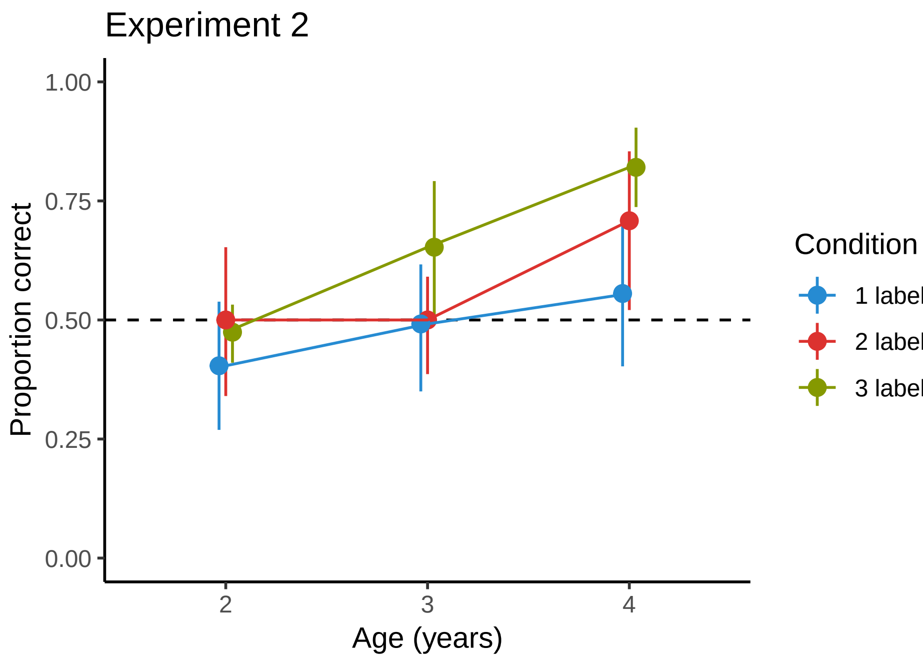

Figure 6. Accuracy data for three age groups across three different conditions. Conditions varied by the number of times the child observed an unambiguous novel label applied to the familiar object prior to the critical disambiguation trial. The dashed line corresponds to chance. Ranges are 95% confidence intervals.

Appendix

Vocabulary Assessment Items (Exp. 1).

1. hatchet
2. elephant
3. flamingo
4. duck
5. hug
6. broccoli
7. panda
8. hexagon
9. parallelogram
10. carpenter
11. drum
12. chef
13. bear
14. harp
15. vase
16. globe
17. triangle
18. vegetable
19. beverage
20. goat

Familiar Words (Exp. 1).

1. bottle

2. cup
3. spoon
4. bowl
5. apple
6. cookie
7. banana
8. pretzel
9. ball
10. shoe
11. flower
12. balloon
13. guitar,
14. bucket

Novel Words (Exp. 1).

1. kettle
2. ladle
3. whisk
4. tongs
5. radish
6. leek
7. bok choy
8. kumquat
9. rudder
10. beaker
11. funnel
12. disk

13. bung
14. cam
15. chestnut
16. dulcimer
17. fig
18. ginger
19. gourd
20. longan
21. luffa
22. okra
23. pipette
24. sieve