The role of developmental change and linguistic experience in the mutual exclusivity effect

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

[1] University of Chicago

[2] University of Wisconsin-Madison

[3] Gallaudet University

[4] New York University

[5] Cognitive ToyBox, Inc.

[6] Stanford University

Author Note

We would like to thank support from Cognitive ToyBox, and note that BML and TK are co-founders of Cognitive ToyBox which developed the app in Experiment 1. Data from Experiment 2 were previously presented in the Proceedings of the Cognitive Science Society Conference in Lewis & Frank (2013).

Correspondence concerning this article should be addressed to Molly Lewis, 5735 S Ellis Ave, Chicago, IL 60637. E-mail: mollyllewis@gmail.com

Abstract

Given a novel word and a familiar and a novel referent, children have a bias to assume the novel word refers to the novel referent. This bias – often referred to as the "Mutual Exclusivity" (ME) effect – is thought to be a potentially powerful route through which children might learn new word meanings, and, consequently, has been the focus of a large amount of empirical study and theorizing. Here, we focus on two aspects of the effect that have received relatively little attention in the literature: Development and experience. In particular, we characterize change in the strength of the effect across development, and investigate the role that linguistic experience plays in this developmental change. We first summarize the current body of empirical findings via a meta-analysis, and then present two experiments that examine the relationship between a child's amount of linguistic experience and the strength of the ME effect. We conclude that the strength of the effect varies dramatically across development and that linguistic experience is likely one causal factor contributing to this change. In the General Discussion, we describe how existing theories of ME can account for our findings, and highlight the value of computational modeling for future theorizing.

Keywords: mutual exclusivity, disambiguation effect, word learning, meta-analysis, linguistic experience, developmental change

Word count: approx. 9200

The role of developmental change and linguistic 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 the current 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 development and experience.

Before engaging with the prior literature related this behavior, it is useful to first make several theoretical distinctions and clarify terminology. 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" in the literature. 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 type of exposure was enough for the child to establish some representation of the link between the phonological form of the novel word 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 the case of 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 ME inference. In this work, children were presented 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. Our work here focuses only on the more narrow question of how children make the inference in the context of the ME paradigm.

Here we adopt the label "mutual exclusivity" (ME) effect as a generic term referring 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 all 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 ME inference would lead her to rule out all meanings for which she already had a name. With this restricted space of possibilities, she is more likely to identify the correct referent than if all objects in the context were considered as candidate 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 turquoise reference situations would further support the inference (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"). Subordinate and superodinate labels present a particular challenge to the learner since 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 of language ("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 regularity in language. Our aim here is not to definitively discriminate between theories of ME, but rather present evidence for a causal role of experience in the ME effect that can provide a constraint on existing theories of ME. 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 effect strength. In part, this focus may be due to methodological challenges in conducting developmental experiments rather than to an underlying theoretical motivation: 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 of 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 (e.g., Bion et al., 2013; Halberda, 2003; W. E. 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 were 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 the 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 contributing 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 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 ME effect as well a moderating factors, we conducted a meta-analysis on the existing body of literature investigating the ME 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 used 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

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

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

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 below remain qualitatively 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. Papers included in the meta-analysis are marked with an asterick in the bibliography.

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, 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), and participant population type.[3] 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 been studied in the literature: (a) typically-developing monolingual children, (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), standard deviations for novel object selections, t-statistic, and Cohen's d, where available. 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 ds), but we were able to estimate the means and standard deviations though measurement of plots ($N = 13$), imputation from

[3]We also coded a number of other moderating variables not included here: method (eyetracking or pointing), number of alternatives in the forced choice task, and task modality (paper vs. object). See http://metalab.stanford.edu/ for these analyses.

other data within the paper ($N = 4$), 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 in R (Viechtbauer & others, 2010) using mixed-effect models with grouping by paper.[4] 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: Familiar-Novel Trials. We first examined effect sizes of ME for typically-developing children in the canonical familiar-novel paradigm. This is the central data point that theories of ME 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 ($Q = 968.13$; $p <.001$).

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

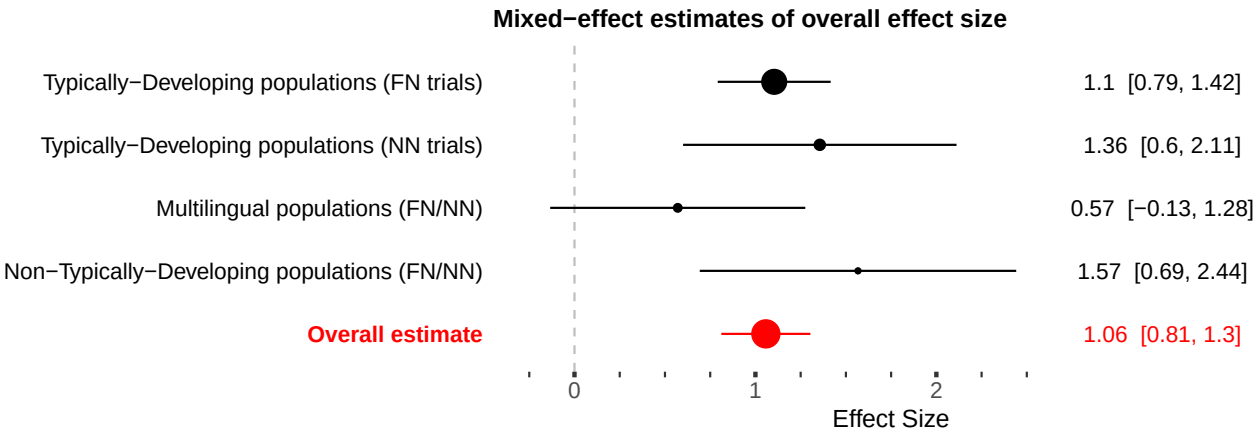

Figure 1. Mixed-effect effect size estimates for all conditions (red) and each of the four theoretically-relevant conditions in our sample. Ranges are 95 percent confidence intervals. Point size corresponds to sample size. FN = Familiar-Novel trials; NN = Novel-Novel trials.

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 was a reliable predictor of effect size ($\beta = 0.05$, $Z = 11.85$, $p <.001$; see Table 1), suggesting that the ME effect becomes larger as children get older (Figure 2). 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 ME effect 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. Age ($\beta = 0.07$, $Z = 2.14$, $p = 0.03$), but not vocabulary size ($\beta = 0.00$, $Z = 0.31$, $p = 0.75$), was a reliable predictor of ME effect size.

These analyses confirm that the ME effect is robust, and associated with a relatively large effect size ($d = 1.1$ [0.79, 1.42]). They also suggest that the magnitude of the effect strengthens over development. Vocabulary size, though correlated with age, does not predict additional effect size variance over and above age. This finding is difficult to interpret however, given the fact that estimates of vocabulary size are likely far less accurate than those of age, and we likely have less power to detect an effect of vocabulary size relative to

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 parameters for four separate conditions. Ranges are 95 percent confidence intervals.

Model	n	term	estimate [CI]	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.

age, since estimates of vocabulary size are available for only a minority of conditions (20%).

Typically-Developing Population: Novel-Novel Trials. One way that vocabulary knowledge could lead to increased performance on the Familiar-Novel ME task is through increased certainty about the label associated with the familiar word: If a child is more certain that a ball is called "ball," then the child should be more 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 ME trial, where this previously-learned label becomes the "familiar" object in the ME task. If knowledge of the familiar object is not the only contributor to

age-related changes in the ME effect, then there should an increase in the magnitude of the ME effect in Novel-Novel trials, as well as Familiar-Novel trials. In addition, if the strength of knowledge of the "familiar" object influences the strength of the ME 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 ($\beta = 0.03$, $Z = 3.55$, p <.001), suggesting that the strength of the ME effect 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 this trial type.

Finally, we fit a model with both age and trial type (Familiar-Novel or Novel-Novel) as moderators of the ME effect. Both moderators predicted independent variance in ME effect size (age: $\beta = 0.04$, Z = -7.05, p <.0001; trial-type: β = -1.08, Z = -7.05, p <.0001), with Familiar-Novel conditions and conditions with older participants tending to have larger effect sizes.

These analyses suggest that both development (either via maturation or experience-related changes) as well as the strength of the familiar word representation are related to the strength of the ME effect. A successful theory of ME 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 ME effect relies on mechanisms that are domain-general and independent of linguistic knowledge, then we should expect the

magnitude of the effect to be the same for multilingual children compared to monolingual children.

Children learning multiple languages reliably showed the ME 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 children with diagnoses of Autism-Spectrum Disorder (ASD), Mental Retardation, Williams Syndrome, Late-Talker, Selective Language Impairment, and deaf/hard-of-hearing. While this sample is highly heterogeneous, we group them together due to the sparsity of data on any single non-typical population. 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 ME 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 the ME task, to a first approximation this information might suggest that social reasoning skills are not critically involved in making ME inferences (de Marchena, Eigsti, Worek, Ono, & Snedeker, 2011; Preissler & Carey, 2005). As a heterogeneous group, these studies can provide evidence about the extent to which the ME behavior is robust to developmental differences.

Overall, non-typically developing children succeeded on the ME task ($d = 1.57$ [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. Population type was a reliable predictor of effect size with non-typically developing children tending to have a smaller bias compared to typically developing children ($\beta = -0.50$, $Z = -2.86$, $p < .0001$). Age was also a reliable predictor of effect size in this model ($\beta = 0.04$, $Z = 11.34$, $p < .0001$).

This analysis suggests that non-typically developing children succeed in the ME paradigm just as typically developing children do, albeit at lower rates, and show the same broad developmental trajectory. Theoretical accounts of ME will need to account for how non-typically developing children are able to develop the ability to make the ME inference, despite a range of different cognitive impairments.

Discussion

To summarize our meta-analytic findings, we find a robust ME 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 Familiar-Novel 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 ME effect. First, the strength of the bias increases across

development, independent of the strength of the learners knowledge of the "familiar" word. 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 the effect in multilingual children (who presumably have less language experience with any particular language). Second, developmental change in the strength of the ME effect is observed for children across a variety of populations, suggesting that developmental change is a robust pattern and is related to the the mechanism underlying the ME effect for different populations.

There is also some evidence that language experience accounts for developmental change on the basis of the fact that we see a larger effect size in Familiar-Novel trials compared to Novel-Novel trials. Nevertheless, the meta-analytic approach is limited in its ability to measure the relationship between linguistic experience and developmental change since few studies in our sample measure vocabulary size ($N = 8$), and even fewer measure vocabulary size at multiple ages within the same study ($N = 4$; Horst, Scott, & Pollard, 2010a; Markman et al., 2003; Mather & Plunkett, 2009a; A. Williams, 2009). In the next section, we use experimental methods to more directly examine the relationship between linguistic experience and the ME effect.

Experiment 1: ME Effect and Vocabulary Size

In Experiment 1, we test the prediction that children with larger vocabularies should show a strong ME effect by measuring vocabulary size in a large sample of children across multiple ages who also completed the ME task. We find that vocabulary size is a strong predictor of the strength of the ME effect across development and that vocabulary size predicts more variance than developmental age.

Table 2

Demographics of children in Experiment 1.

Age group	Mean age (months)	Sample size
2-yo	30.02	69
3-yo	41.64	85

Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Our hypotheses and analysis plan were pre-registered (https://osf.io/tt29f/register/5771ca429ad5a1020de2872e), and we note below where our analyses diverged from this pre-registration.

Participants. We conducted a range of power analyses to determine our sample size and found that we needed a large sample size to estimate the unique effect of vocabulary on accuracy, since vocabulary and age tend to be highly correlated with each other. We registered our target sample size to be 80 2-year-olds and 80 3-year-olds. In total, 172 children completed the task (2-yo: $N = 79$; 3-yo: $N = 93$). We excluded participants who did not correctly answer at least half of the familiar noun control trials ($N = 9$), as described in our pre-registration. In addition, there were 9 children in our sample whose parents reported that they were exposed to English less than 75% of the time. We excluded these participants from our analysis since previous work has suggested that the ME effect is affected by multilingualism (e.g., Byers-Heinlein & Werker, 2009), though note that this a conservative choice since our meta-analysis did not reveal an effect of multilingualism on the ME effect. Exclusions on the basis of language input were not described our pre-registration analysis plan, but all analyses remain qualitatively the same when these children are included in the sample. Our final sample included 154 children ($N_{female} = 93$; see Table 2).

Stimuli. The ME 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 1-3 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 (PPVT; see Appendix; L. M. Dunn, Dunn, Bulheller, & Häcker, 1965). We selected words for our vocabulary assessment on the basis of pilot testing and age of acquisition data from the Wordbank database (M. C. Frank, Braginsky, Yurovsky, & Marchman, 2017) with the goal of identifying words that would be challenging for children across the target age range. We developed our own very short, tablet-based assessment of vocabulary size because the complete PPVT would be prohibitively time consuming and the CDI could not be used with our full target age range.

Design and Procedure. In order to test a large sample of children, we designed a short and simple testing procedure that could be conducted on a tablet in a museum setting. 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. 3). 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.

In the test phase, each test trial consisted of two screens: One presenting a single object and an unambiguous label (Fig. 3b), and another presenting two objects and a single label (Fig. 3c). 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 ME 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 leek). 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 3). Trials were presented randomly, and children were only allowed to make one selection.

Table 3

Design for each of the four trial types. "N" indicates a novel referent and "F" indicates a familiar referent. Each test trial involved two displays. The first display introduced an object and its label unambiguously; 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

After the ME 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 ME task were coded as correct if the participant selected the familiar object on Control trials and the novel object on Experimental 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). All ranges are 95% confidence intervals. Effect sizes are Cohen's d values.

Results

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.[5] There was a main effect of trial type, suggesting that participants were less accurate in NN trials compared to NF trials (β = -0.87, SE = 0.2, Z = -4.4, $p < .001$). There was also a marginal main effect of target type, with novel referents being more difficult for children than familiar referents (β = -0.48, SE = 0.24, Z = -1.99, p = 0.05). Finally, there was a marginal interaction between the two factors (β = 0.38, SE = 0.24, Z = 1.61, p = 0.11), 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

[5]The model specification was as follows: `accuracy ~ target.type * trial.type + (target.type | subject.id) + (trial.type | subject.id)`.

Table 4

Parameters of logistic mixed model predicting accuracy on ME 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)	2.00	0.15	12.94	<.0001
Vocabulary	6.12	1.06	5.77	<.0001
Trial Type (NN)	-0.34	0.24	-1.46	0.14
Age	0.01	0.02	0.67	0.51
Vocabulary x Trial Type (NN)	-2.56	1.52	-1.68	0.09
Vocabulary x Age	-0.01	0.14	-0.06	0.96
Age x Trial Type (NN)	0.02	0.03	0.57	0.57
Vocabulary x Age x Trial Type (NN)	0.17	0.20	0.84	0.4

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 4 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).[6] Table 4 presents the model parameters. The only reliable predictor of accuracy was vocabulary size

[6]The model specification was as follows: `accuracy ~ vocabulary.size * age * trial.type + (trial.type | subject.id)`

(β = 6.12, SE = 1.06, Z = 5.77, p <.0001), suggesting that children with larger vocabularies tended to be more accurate in the ME task. Vocabulary size did not interact with trial type (β = -2.56, SE = 1.52, Z = -1.68, p = 0.09), suggesting that children with larger vocabularies were more likely to make the ME inference in both NF and NN trials. Notably, age was not a reliable predictor of accuracy over and above vocabulary size (β = 0.01, SE = 0.02, Z = 0.67, p = 0.51).

Discussion. Experiment 1 examines the relationship between the strength of the ME effect and vocabulary size. We find that the strength of the ME effect is highly predicted by vocabulary size, with children with larger vocabularies tending to show a larger ME effect. 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 5 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.[7] 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. The experimental data thus provide converging evidence with the meta-analysis that there is developmental change in the strength of the bias, and that the effect is weaker for NN trials.

There are, however, some notable differences between the Experiment 1 data and the meta-analytic results. First, while the direction of the influence of age on the ME effect is the same in both studies, the magnitude of the developmental effect is much smaller in Experiment 1 relative to the meta-analytic data within the same 24- to 48- month developmental range. This difference could be due to the fact that researchers in the

[7]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.*

meta-analytic studies calibrate their method to the age of their participants (e.g., eye-tracking for younger children and pointing for older children), and there is evidence that the effect size of a method varies across development (Bergmann et al., 2018). Second, the variance is larger for the meta-analytic estimates compared to the experimental data, presumably because there is more heterogeneity across experiments than across participants within the same experiment. Third, the magnitude of the effect of trial type (NF vs. NN) is much smaller in the experimental data, relative to the meta-analytic data. This incongruence could be due to any number of differences across studies, such as the difficulty of the familiar word in NF paradigms.

In addition, the data from Experiment 1 provide new evidence relevant to the mechanism underlying the effect: children with larger vocabulary tend to have a stronger ME bias. In principle there are two ways that vocabulary knowledge could support the ME 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. The relationship between vocabulary size and the magnitude of the effect in NN trials 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.

Regardless about the specific route through which vocabulary knowledge influences the ME inference, the hypothesized relationship between experience and the ME effect is fundamentally causal. Nevertheless, the data from both the meta-analytic study and the current experiment only provide correlational evidence about their relationship. In Experiment 2, we aimed to more directly test the causal hypothesis by experimentally

Table 5

Demographics of children in Experiment 2.

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

manipulating the strength of the learner's knowledge about the familiar object.

Experiment 2: Mutual Exclusivity Effect and Familiarity

In Experiment 2, we used the same design as in the Novel-Novel trials from Experiment 1, but manipulated the amount of exposure children were given to the novel object and label prior to the critical ME trial. We reasoned that children who observed more instances of a novel label referring to a novel object should have higher certainty about the label name. If the strength of knowledge about the "familiar" object influences the strength of the ME effect, then we should expect a larger ME effect when 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 5). 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.

Stimuli. The referent objects were the set of 8 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). All items were distinct in color and shape. The novel words were 8 single syllable labels (e.g., "dax," "zot," and "gup").

Design and Procedure. Each child completed four trials. Each trial consisted of a training and a test phase in a "novel-novel" ME 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. 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 6

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

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

Children showed a stronger ME effect with development and as the number of training labels increased (Figure 6).

We analyzed the results using a logistic mixed model to predict correct responses with age, number of labels, and their interaction as fixed effects.[8] Model results are shown in Table 6. There was a significant effect of both age and number of labels: Children who were older and observed the occurrences of label for the "familiar" object showed stronger ME effect. The interaction between age and number of labels was not significant.

Experiment 1 thus provides causal evidence for a link between the strength of knowledge about the "familiar" word in the ME task and the strength of the ME effect: A stronger representation about the "familiar" word in the ME task leads to a stronger ME inference. This pattern of findings is consistent with the correlational relationship observed in Experiment 1 in which children with larger vocabularies tended to show a larger ME

[8]The model specification was as follows: `accuracy ~ times.labeled * age + (times.labeled | subject.id)`.

effect. We cannot, however, compare the magnitude of the effects in the two experiments since a few exposures to a novel label in the laboratory is not straight-forwardly comparable to the history of labeling experiences that a child encounters in their natural environment. Nevertheless, Experiment 2 provides strong evidence that at least part of the relationship between vocabulary size and the strength of the ME effect observed in Experiment 1 is due to children's knowledge about the familiar object label.

General Discussion

We set out to measure developmental and experience-based shifts in children's ability to make ME inferences. Across a systematic meta-analysis of the existing literature and two new studies, we found strong evidence that older children make stronger and more reliable ME inferences than younger children. Further, both the meta-analytic findings and the results of Experiment 1 suggest that vocabulary size is related to ME performance, perhaps moreso than age. Finally, Experiment 2 showed that ME inference strength is also directly influenced by children's familiarity with the alternative objects and their labels. Taken together, this body of evidence suggests that the ability to make ME inferences changes very substantially with development and experience, changes that have been under-appreciated due to the limited size and developmental range of most of the studies of this phenomenon.

The role of development in theories of the ME effect

We next turn to the implications of these findings for theories of ME. The literature contains a large number of proposals for the mechanisms supporting ME, and many of these overlap or differ only in subtle ways. Here we briefly describe several influential proposals, highlighting the commonalities and differences across theoretical views and considering the ways they could accommodate our findings. To summarise our conclusion, developmental

and experience-based changes in the strength of the ME inference are not *inconsistent* with many possible theoretical alternatives in the sense that there are not clear predictions that a specific ability would *not* develop. Instead, most theories simply have not discussed the predicted developmental course of the ME inference explicitly; developmental and experience-based change are auxiliary to the theory. In contrast, computational models of word learning – as learning models – make clear and explicit predictions about the role of experience. Given this, our work here suggests that such models may provide a more parsimonious framework for thinking about ME.

Constraint and bias accounts. One influential proposal regarding the sources of ME inferences is that children have a constraint that is innate or early-emerging. 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, & Nisbett, 1973).

This classic constraint-based theory of ME does not have an obvious role for the developmental and experiential effects we have documented here. Since even young children are posited to have some bias, on such a theory, developmental effects on this kind of theory would be primarily generated by changes in downstream, performance-based factors (for

example, the ability to attend to the experimental task). Further, experience-based effects such as those observed in our Experiment 2 could be the result of individual children simply failing to access individual lexical representations. These modifications to constraint theories are ad-hoc (but perhaps not unreasonable).

Another related constraint-based proposal is the Novel-Name Nameless-Category principle (N3C; Golinkoff et al., 1994; Mervis & Bertrand, 1994). On the N3C account, 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). Unlike the ME constraint, the N3C principle was argued (based on the empirical finding of developmental change) to emerge developmentally with language experience. Nevertheless, the specific developmental prediction was that N3C became available after children went through a "vocabulary spurt" rather than emerging gradually and continuing to increase in strength (as we observed). Further, this account does not have an account of specific experiences with particular words, the effect we observed Experiment 2.

Pragmatic contrast accounts. One important alternative to principle-based accounts are pragmatic accounts. Under these accounts, the ME inference derives from reasoning about the intention of the speaker within the current 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).

Although developmental and experience-based effects were not a specific focus of these accounts, these findings are relatively easy to accommodate within this framework. A pragmatic theorist could simply argue that children's understanding of each of these principles is changing across the relevant time period (Clark & Amaral, 2010). Experiential effects similarly are not accounted for in this framework, but could be added as an auxiliary assumption.

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.

Although this proposal was formulated on the basis of developmental data showing failures at 14 months (with an interesting pattern of alternative behavior), there is no account provided for what sorts of developmental changes or experiences lead to the emergence of disjunctive syllogism. Indeed, syllogistic reasoning of this sort is argued to be available even in younger children (Cesana-Arlotti et al., 2018; Halberda, 2018). If so, again, auxiliary theoretical assumptions are required to specify the specific maturational processes or developmental experiences that lead the inference to become available for older children.

Probabilistic accounts. Probabilistic computational accounts contend that ME does not derive from an explicit representation of a constraint or principle nor from pragmatic reasoning, as proposed by other accounts. Rather, under this broad class of accounts accounts, the ME inference is the product of a word learning system that tracks the frequency of words and their referents over time, and then uses probabilistic associative mechanisms to infer novel word-referent mappings.

There are a wide variety of computational models that instantiate such ideas. For example, in an early model Regier (2005) used an associative exemplar model to account for a variety of influential findings in early word learning including the ME inference. Under this model, second labels are hard to learn due to memory interference (and hence novel labels are preferentially mapped to new referents). Similarly, in the model of Frank, Goodman, and Tenenbaum (2009), a set of simple parsimony biases lead the model to assume that it is more likely that a novel word would have been used to refer to a novel referent (rather than a familiar word also having a second meaning that was never used). While the details vary for other models, the general set of principles in operation is similar in models by e.g., McMurray, Horst, and Samuelson (2012), Fazly, Alishahi, and Stevenson (2010), and Kachergis, Yu, and Shiffrin (2012).

Unlike the largely verbal theories described above, these computational models allow the evaluation of both developmental and experiential effects. In fact, the findings of our meta-analysis and Experiments 1 and 2 should emerge in some form from nearly all of the computational models mentioned above. For example, the strength of the ME inference in the model of Frank et al. (2009) is directly proportional to the number of observations of the familiar word. Thus, more experience with language will lead to more robust representations of familiar words and stronger ME inferences. Similarly, within the framework of Experiment 2, the number of experiences with the first novel word should mediate the strength of the inference to the second (this finding is demonstrated through simulation in a related model

by Lewis & Frank, 2013). In general, these computational models posit that ME inferences emerge from computations over graded representations. These representations could be graded, memory representations (Kachergis et al., 2012; Regier, 2005) or neural network weights (McMurray et al., 2012); they could also be probabilities in a more explicit representation of the lexicon (Fazly et al., 2010; Frank et al., 2009).

The broader point is that, on most of the verbal theories described above, developmental and experience-based changes in ME are auxiliary to the core theory of the phenomenon. Even those theories that have some role for development only discuss the notion of developmental emergence based on a linguistic generalization or a vocabulary milestone. In contrast, each of these computational theories is a learning theory: it takes experience with a particular stimulus as a core part of the theory. Thus, our findings are much more clearly captured by the computational literature on modeling early word learning than by the verbal theories that preceded it.

The next step in this literature – one that we hope is provoked by our work – is to explore quantitative fits to specific developmental patterns. While all of the models described above can in principle provide quantitative predictions, in practice it will take significant work to create a fair comparison of the shape of these predictions to the trends we observed here. Such quantitative modeling of developmental change would provide a powerful step forward in terms of using insights from the literature to predict variation amongst children.

Conclusions

Our theorizing about word learning has often taken as the primary phenomenon the emergence of a particular phenomenon. The associated theorizing then often provides only a relatively small part for further developmental change, if any at all. Similarly, while no theorist would deny the importance of experience with a particular stimulus as moderating a

specific experimental effect, these experiences are rarely core to the theory being developed. In contrast, in our survey of the literature and our experiments, we found that both experience and development were key quantitative determinants of children's ability to perform the ME inference. Thus, such models provide a parsimonious starting point for reasoning about the origins of ME. Further, and more broadly, the development of explicit computational theories provides a route to incorporate developmental experience more explicitly into our theorizing.

References

References marked with an asterisk indicate studies included in the meta-analysis.

* Allen, R., & Scofield, J. (2010). Word learning from videos: More evidence from 2-year-olds. *Infant and Child Development*, *19*(6), 649–661. https://doi.org/10.1002/icd.712

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

* Bedford, R., Gliga, T., Frame, K., Hudry, K., Chandler, S., Johnson, M. H., ... others. (2013). Failure to learn from feedback underlies word learning difficulties in toddlers at risk for autism. *Journal of Child Language*, *40*(1), 29–46.

* Behrend, J. S. A. D. A. (2007). Two-year-olds differentially disambiguate novel words and facts. *Journal of Child Language*, *34*(04). https://doi.org/10.1017/s0305000907008100

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*, 1996–2009.

* Beverly, B., & Estis, J. (2003). Fast mapping deficits during disambiguation in children with specific language impairment. *Journal of Speech Language Pathology and Audiology*, *27*(3), 163–171.

* 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.

* 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. https://doi.org/10.1016/j.cognition.2013.05.010

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

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

Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., & Bonatti, L. L. (2018). Precursors of logical reasoning in preverbal human infants. *Science*, *359*(6381), 1263–1266.

* Choi, I.-R., & Hwang, M. (2014). Korean late-talkers' use of the mutual exclusivity assumption on first versus second label learning. *Communication Sciences & Disorders*, *19*(3), 285–293.

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.

Clark, E. V., & Amaral, P. M. (2010). Children build on pragmatic information in language

acquisition. *Language and Linguistics Compass*, *4*(7), 445–457.

* Davidson, D., Jergovic, D., Imami, Z., & Theodos, V. (1997). Monolingual and bilingual children's use of the mutual exclusivity constraint. *Journal of Child Language*, *24*(1), 3–24. https://doi.org/10.1017/s0305000996002917

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.

* Deák, G. O., Yen, L., & Pettit, J. (2001). By any other name: When will preschoolers produce several labels for a referent? *Journal of Child Language*, *28*(03). https://doi.org/10.1017/s0305000901004858

* 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.

* Estis, J. M., & Beverly, B. L. (2015). Children with sli exhibit delays resolving ambiguous reference. *Journal of Child Language*, *42*(1), 180–195.

* Evey, J. A., & Merriman, W. E. (1998). The prevalence and the weakness of an early name mapping preference. *Journal of Child Language*, *25*(1), 121–147. https://doi.org/10.1017/s030500099700336x

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, I. (1999). The use of word-learning principles in young monolingual and bilingual children. *Doctoral Dissertation*.

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

* Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development*, *17*(1), 1–17. https://doi.org/10.1080/15248372.2015.1061528

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., 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.

* Gollek, C., & Doherty, M. J. (2016). Metacognitive developments in word learning: Mutual exclusivity and theory of mind. *Journal of Experimental Child Psychology*,

148, 51–69. https://doi.org/10.1016/j.jecp.2016.03.007

* Graham, S. A., Nilsen, E. S., Collins, S., & Olineck, K. (2010). The role of gaze direction and mutual exclusivity in guiding 24-month-olds' word mappings. *British Journal of Developmental Psychology*, *28*(2), 449–465.

* Graham, S. A., Poulin-Dubois, D., & Baker, R. K. (1998). Infants disambiguation of novel object words. *First Language*, *18*(53), 149–164. https://doi.org/10.1177/014272379801805302

* Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science*, *13*(1), 252–263. https://doi.org/10.1111/j.1467-7687.2009.00871.x

* 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*. https://doi.org/10.3389/fpsyg.2015.01200

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

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

* Halberda, J. (2006). Is this a dax which i see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognitive Psychology*, *53*(4), 310–344. https://doi.org/10.1016/j.cogpsych.2006.04.003

Halberda, J. (2018). Logic in babies. *Science*, *359*(6381), 1214–1215.

* Hollich, G., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R., Brown, E., Chung, H., ... 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.

* 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., Scott, E. J., & Pollard, J. A. (2010). The role of competition in word learning via referent selection. *Developmental Science*, *13*(5), 706–713. https://doi.org/10.1111/j.1467-7687.2009.00926.x

* Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy*, *15*(2), 125–150. https://doi.org/10.1111/j.1532-7078.2009.00009.x

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

* Jarvis, L. H., Merriman, W. E., Barnett, M., Hanba, J., & Van Haitsma, K. S. (2004). Input that contradicts young children's strategy for mapping novel words affects their phonological and semantic interpretation of other novel words. *Journal of Speech, Language, and Hearing Research*, *47*(2), 392–406.

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.

* Kalashnikova, M., Mattock, K., & Monaghan, P. (2016a). Flexible use of mutual exclusivity in word learning. *Language Learning and Development*, *12*(1), 79–91.

* Kalashnikova, M., Mattock, K., & Monaghan, P. (2016b). Mutual exclusivity develops as a consequence of abstract rather than particular vocabulary knowledge. *First Language*, *36*(5), 451–464. https://doi.org/10.1177/0142723716648850

* Lederberg, A. R., Prezbindowski, A. K., & Spencer, P. E. (2000). Word-learning skills of

deaf preschoolers: The development of novel mapping and rapid word-learning strategies. *Child Development*, *71*(6), 1571–1585. https://doi.org/10.1111/1467-8624.00249

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. L., & 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* (Vol. 35).

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. https://doi.org/10.1080/15250000802569702

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., & Marazita, J. M. (1995). The effect of hearing similar-sounding words on young 2-year-olds disambiguation of novel reference. *Developmental Psychology*, *31*(6), 973–984. https://doi.org/10.1037/0012-1649.31.6.973

* Merriman, W. E., & Schuster, J. M. (1991). Young children's disambiguation of object name reference. *Child Development*, *62*(6), 1288. https://doi.org/10.2307/1130807

* 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.

* Merriman, W. E., Marazita, J., & Jarvis, L. H. (1993). Four-year-olds' disambiguation of action and object word reference. *Journal of Experimental Child Psychology*, *56*(3), 412–430. https://doi.org/10.1006/jecp.1993.1042

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

* Mervis, C. B., & Bertrand, J. (1995). Acquisition of the novel name-nameless category (N3C) principle by young children who have down syndrome. *American Journal on Mental Retardation*.

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.

* Scofield, J., & Williams, A. (2009). Do 2-year-olds disambiguate and extend words learned from video? *First Language*, *29*(2), 228–240. https://doi.org/10.1177/0142723708101681

* Sugimura, T., & Sato, N. (1996). Factors affecting assumptions about mutual exclusivity and novel name-nameless category. *Perceptual and Motor Skills*, *82*(3_suppl), 1147–1153. https://doi.org/10.2466/pms.1996.82.3c.1147

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.

* Wall, J. L., Merriman, W. E., & Scofield, J. (2015). Young children's disambiguation

across the senses. *Cognitive Development*, *35*, 163–177. https://doi.org/10.1016/j.cogdev.2015.06.001

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

* White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, *59*(1), 114–132. https://doi.org/10.1016/j.jml.2008.03.001

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

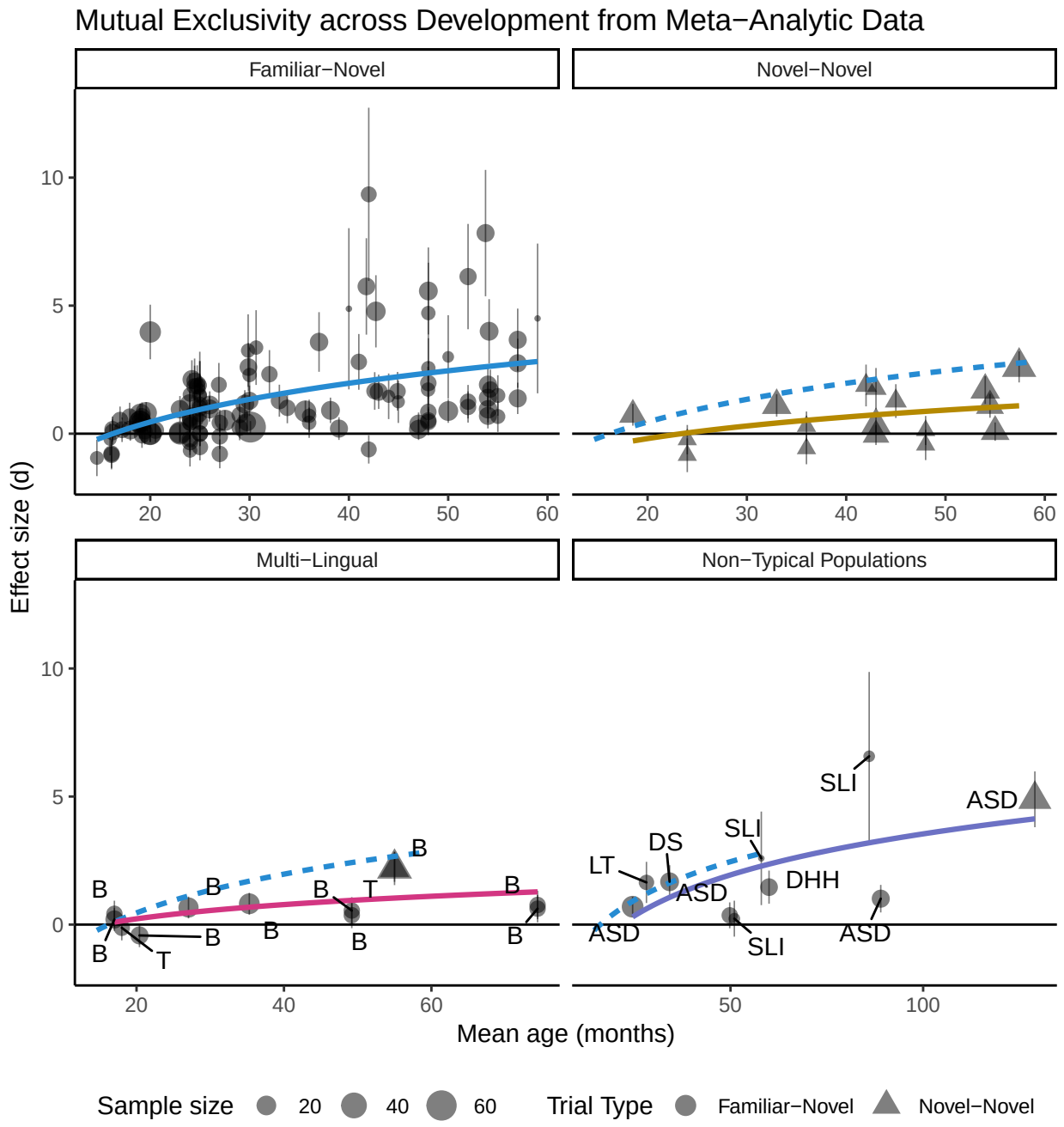

Figure 2. Developmental trajectory of ME effect as a function of four moderators tested in the meta-analysis. For reference, the model fit for familiar-novel trials in typical populations is shown on each moderator plot (blue dash). Ranges correspond to 95% confidence intervals. Model fits shown here 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 impairement; DHH = deaf/heard-of-hearing.

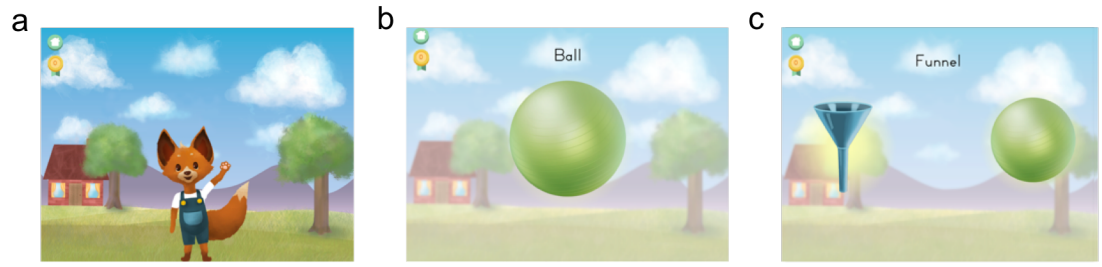

Figure 3. Example screenshots for a Experimental Novel-Familiar test trial in Experiment 1. On each test trial, Mr. Fox first appeared to get the child's attention (a). Next, an object appeared and was labeled through the tablet speakers ('It's a ball'; b). Two objects then appeared and children were asked to make a selection ('Touch the funnel'; c).

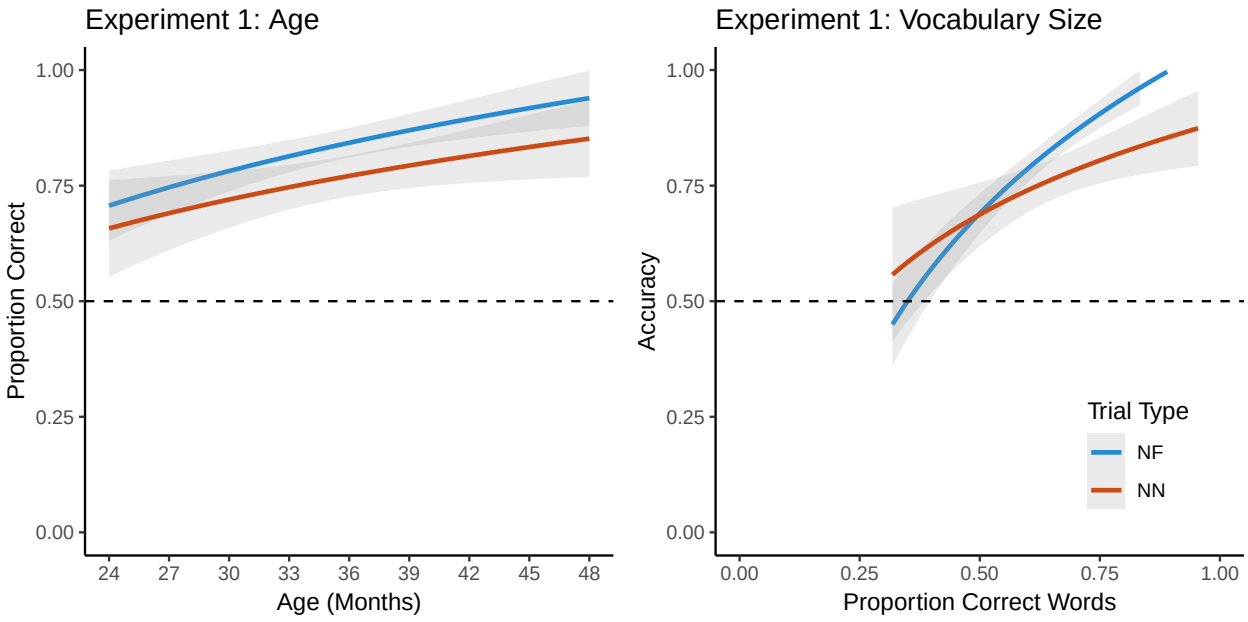

Figure 4. Experiment 1 results. Accuracy as a function of age (months; left) and vocabulary size (proportion correct on vocabulary assessment; right). Blue corresponds to trials with the canonical novel-familiar ME task, and red corresponds to trials with two novel alternatives, where a novel of label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

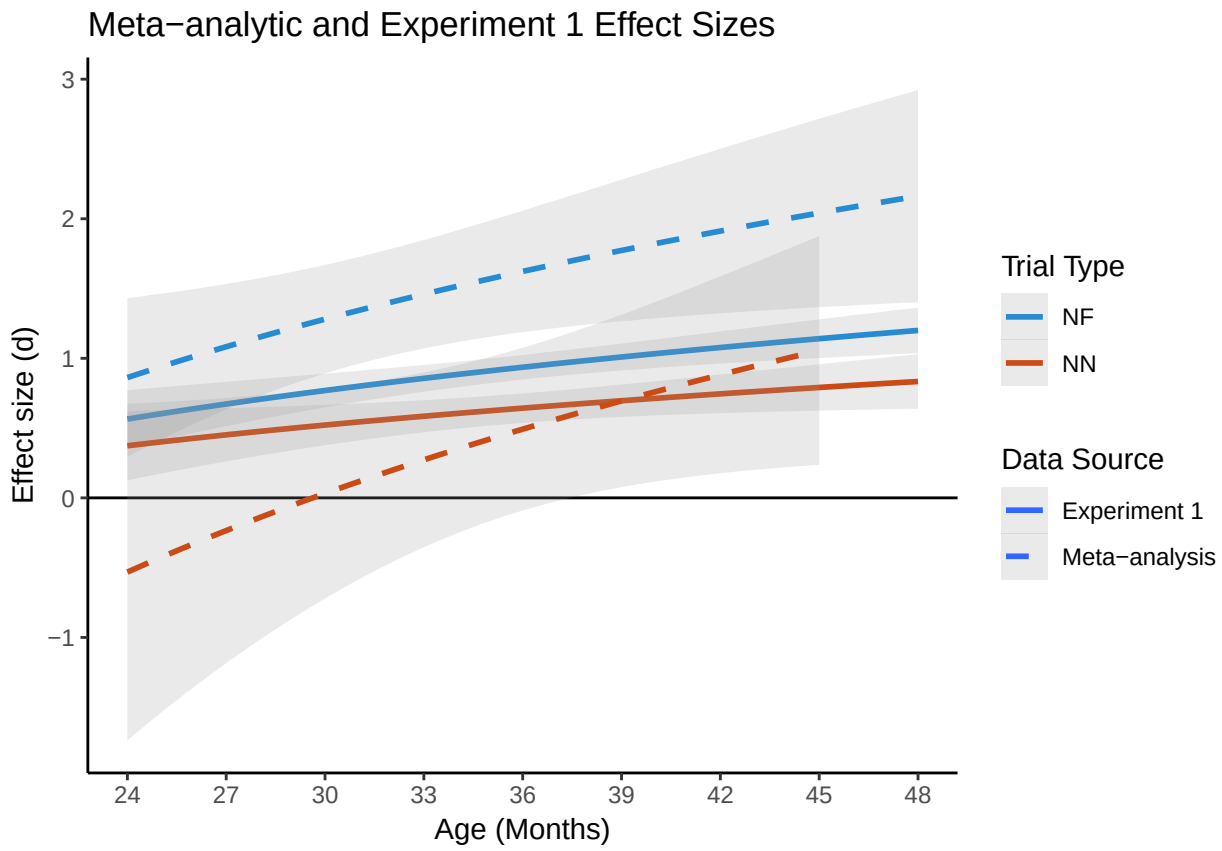

Figure 5. Meta-analytic data and data from experimental trials in Experiment 1 as a function of age. Effect sizes for Experiment 1 data are calculated for each participant, assuming the across-participant mean standard deviation as an estimate of the participant level standard deviation. Ranges are 95%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

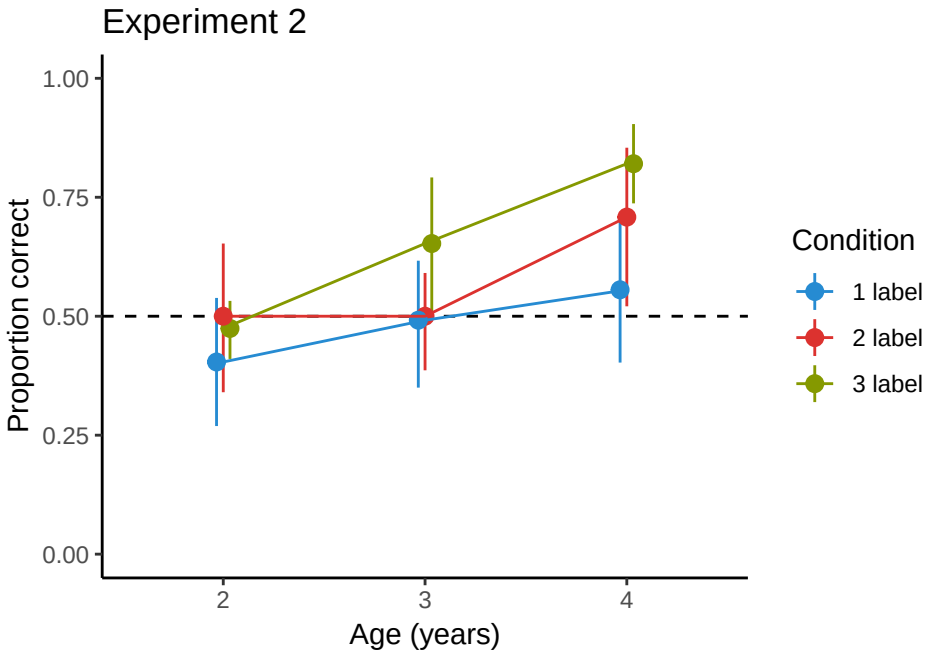

Figure 6. Accuracy data for three age groups across three different conditions in Experiment 2. Conditions varied by the number of times the child observed an unambiguous novel label applied to the familiar object prior to the critical ME trial. The dashed line corresponds to chance. Ranges are 95% confidence intervals. Points jittered along the x-axis for visibility.

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