The role of experience in disambiguation during early word learning

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

Author Note

Data from Experiment 2 were previously presented in the Proceedings of the Cognitive Science Society Conference in Lewis & Frank (2013). *To whom correspondence should be addressed. E-mail: mollylewis@uchicago.edu

Abstract

Young children tend to map novel words to novel objects even in the presence of familiar competitors, a finding that has been dubbed the "disambiguation" effect. This phenomenon is important because it could provide a strong constraint for children in learning new words. But, although the effect is highly robust and widely studied, the cognitive mechanisms underlying it remain unclear. Existing theoretical accounts include a proposal for initial constraints on children's lexicons (e.g. a principle of mutual exclusivity), situation-specific pragmatic inferences, probabilistic accounts, and overhypothesis account. In the current paper, we have two goal: synthesize the existing body of literature and directly examine the causal role of experience on the efect. We present a synthesis of exisiting evidence through a meta-analysis of the existing literature, followed by two experiments that examine the relationship between vocabulary development and the effect. We conclude by summarizing the empirical landscape, and suggest that multiple mechanisms may underly the effect.

Keywords: mutual exclusivity, disambuguation effect, word learning, meta-analysis

Word count: X

The role of experience in disambiguation during early word learning

Introduction

A central property of language is that each word in the lexicon maps to a unique concept, and each concept maps to a unique word (Clark, 1987). Like other important regularities in language (e.g., grammatical categories), children cannot directly observe this general property. Instead, they must learn to use language in a way that is consistent with the generalization on the basis of evidence about only specific word-object pairs.

Even very young children behave in a way that is consistent with this one-to-one regularity in language. Evidence for this claim comes from what is known as the "disambiguation" or "mutual exclusivity" (ME) effect (we return to the issue of nomenclaturer below). 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 in this task tend to choose the novel object as the referent, behaving in a way that is consistent with the one-to-one word-concept regularity in language across a wide range of ages and experimental paradigms (Bion, Borovsky, & Fernald, 2012; R.M. Golinkoff, Mervis, Hirsh-Pasek, & others, 1994; J. Halberda, 2003; Markman, Wasow, & Hansen, 2003; Mervis, Golinkoff, & Bertrand, 1994).

This 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. Critically, the ability to infer that a novel word maps to a novel object makes the problem much easier to solve. 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 child's ability to correctly disambiguate would lead her to rule out all meanings for which she already had a name. With this restricted hypothesis space, the child is more likely to identify the correct referent than if all objects in the context were considered as possible referents.

Despite – or perhaps due to – the attention that the ME effect 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 ("bias accounts," see below), 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? The goal of the current manuscript is to lay out these possibilities and discuss the state of the evidence. Along the way we present a meta-analysis of the extant empirical literature. We then present two new, relatively large-sample developmental experiments that investigate the dependence of children's ME inferences on vocabulary (Experiment 1) and experience with particular words (Experiment 2). We end by discussing the emergence of ME inferences in a range of computational models of word learning. 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.
4. ME inferences are distinct from learning.

A note on 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 RM 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 and effect (selecting the novel object as the referebnt of the novel word) as well as the theoretical account (an early assumption or bias). This conflation of paradigm/effect with theory is problematic, as other authors who have argued against the theoretical account then are in the awkward position of rejecting the name for the paradigm they have used. Other labels (e.g. "disambiguation" or "referent selection" effect) are not ideal, however, because they are not as specific do not refer as closely to the previous literature. Here we adopt the label "mutual exclusivity" (ME) for the general family of paradigms and associated effects, *without* prejudgment of the theoretical account of these effects.

ME has also been referred to as "fast mapping." This conflation is confusing at best. In an early study, S. 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. L. 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 (S. 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 (S. Carey & Bartlett, 1978) or pre-exposure to the object (L. Markson & Bloom, 1997).

Theoretical views of "mutual exclusivity"

What are the cognitive processes underlying this effect? A range of proposals in the literature.

Constraint and bias accounts.

Under one proposal, Markman and colleagues (Markman & Wachtel, 1988,Markman et al. (2003)) suggest that children have a constraint on the types of lexicons considered when learning the meaning of a new word – a "mutual exclusivity constraint." With this constraint, children are biased to consider only those lexicons that have a one-to-one mapping between words and objects. Importantly, this constraint can be overcome in cases where it is incorrect (e.g. property names), but it nonetheless serves to restrict the set of lexicons initially entertained when learning the meaning of a novel word. Under this view, then, the disambiguation effect emerges from a general constraint on the structure of lexicons. This constraint is assumed to be innate or early emerging.

N3C

Probabilistic accounts.

Regier

McMurray

Frank Goodman Tenenbaum

Fazly

Over-hypothesis accounts.

Lewis & Frank (2013)

Pragmatic accounts.

The disambiguation effect is argued to result from online inferences made within the referential context (Clark, 1987; Diesendruck & Markson, 2001). In particular, Clark suggests that the disambiguation effect is due to two pragmatic assumptions held by speakers. The first assumption is that speakers within the same speech community use the same words to refer to the same objects ("Principle of Conventionality"). The second assumption is that different linguistic forms refer to different meanings ("Principle of Contrast"). In the disambiguation task described above, then, children might reason (implicitly) as follows: 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. Thus, under this account, the disambiguation effect emerges not from a higher-order constraint on the structure of lexicons, but instead from in-the-moment inferences using general pragmatic principles.

These two proposals have traditionally been viewed as competing explanations of the disambiguation effect. Research in this area has consequently focused on identifying empirical tests that can distinguish between these two theories. For example, Diesendruck and Markson (2001) compare performance on a disambiguation task when children are told a novel fact about an object relative to a novel referential label. They found that children disambiguated in both conditions and argued on grounds of parsimony that the same pragmatic mechanism was likely to be responsible for both inferences. More recent evidence contradicts this view: tests of children with autism, who are known to have impairments in pragmatic reasoning find comparable performance on the disambiguation task between typically developing children and children with autism (de Marchena, Eigsti, Worek, Ono, & Snedeker, 2011; Preissler & Carey, 2005). This result provides some evidence for the view that disambiguation is due to a domain-specific lexical constraint.

Clark?

In the moment

Learned pragmatics

Logical inference accounts. Justin Halberda (2003)

Theory-constraining findings

NN vs. NF

Speaker-change studies

Autism

Bilingualism

Fast mapping + no retention

Developmental change (halberda)

Synthesis

These are definitely features of a successful account: Timescales - must be one "in the moment" - and one longer-term learned mechanism

Experience

Probabilistic representations

Could be the case also that it's a mixture of pragmatic, etc.

We suggest this competing-alternatives approach to the disambiguation effect should be reconsidered. In a disambiguation task, 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. Both of these constraints would then support children's inferences. In other words, these two classes of theories may be describing distinct, complimentary mechanisms that each contribute to a single empirical phenomenon with their weights in any given task determined by children's age and language experience, the nature of the pragmatic situation, and other task-specific factors.

The current study

Gather evidence on strength of finding

Test emergent relationship to vocabulary (E1)

Test causal relationship to representation strength (E2)

Re-evaluate

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 examines the disamboguation 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). 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 unambigous 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-Familar-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)[1], (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 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 vocabulary size 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

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

number of alternatives in the forced choice task. We used production vocbulary as our estimate of vocabulary size since it was available for more studies in our sample. We coded participant population as one of three subpopulationns 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 imparement, language delays, hearing imparement, 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 ds), but we were able to esimtate 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.[2] In models with moderators, moderators variables were included as additive fixed effects. All estimate ranges are 95% confidence intervals.

[2]The exact model specification was as follows: $model < -metafor :: rma.mv(yi = effect_size, V = effect_size_var, random = \ 1|paper, data = d)$.

Meta-analytic 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.

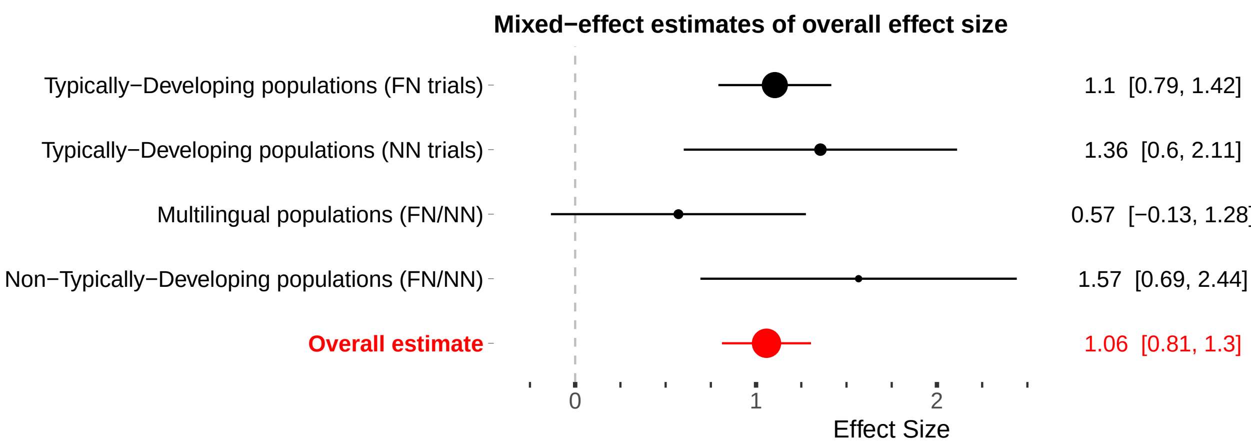


Figure 1. Mixed-effect effect size estimates for all conditions (red) and each of the four theoretically-relevant conditions in our sample. Ranges are 95% confidence intervals. Point size corresponds to sample size. FN = Familiar-Novel trials; NN = Novel-Novel trials.

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

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

We next tried to predict this heterogenity 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 (β = 0.05, z = 11.85, p <.001; see Table X), 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 = 0.03, was a reliable predictor of disambiguation effect size. ACTUALLY THIS ISN'T TRUE (true only for full model)

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 mechansims 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 pure maturational development.

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 "familar" 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 disamiguation effect, then the overall effect size should be smaller for Novel-Novel trials, compared to Familar-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.

Finally, we fit a model with both age and trial type (Familar-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 popuIuation 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 controling 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).

These data do not provide strong evidence that language-specific knoweldge 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 Imparment, and deaf/hard-of-hearing These populations are of theoretical interests because they allow us to observe how imparement to a particular aspect of cognition influences the magnitutude of the disambiguation effect. For example, children with ASD are thought to have impared social reasoning skills (e.g., Phillips, Baron-Cohen, & Rutter, 1998); thus, if children with ASD are able to succeed on disambiguation tasks, this suggests that social reasoning skills are not necessary to making a disambiguation inference.

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 (β = 0.04, z = 3.15, p <.001).

We also asked whether the effect size for non-typically developing children differed from typically-developing children, controling 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, allbeit 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 impairements.

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-Familar paradigm compared to the Novel-Novel paradigm, but both designs show roughly the same developmental trjectory.

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 knolwedge, 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 imparement to effect in multilingual children (who presumably have less language experience with any particular language). Third, children with a range of different imparements are able to make the inference, suggesting that no single mechanism is both necessary and sufficient for the effect.

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 overypothesis 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 imparements suggests that accounts that rely on a single mechanism, such as pragmatic reasoning or a mutual exclusivity constraint alone, are unlikely to describe the mechanim underlying the disambiguation effect across all children.

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 examinine the relationship between linguistic experience and the disambiguation effect.

Experiment 1: ME and Vocabulary

Our meta-analysis points to a robust developmental increase in the strength of the disambiguation effect with age. While all four accounts are able to predict this change, only the overhypothesis account predicts that this increase should be directly related to vocabulary knowledge. 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 multipe 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 predicts more variance than developmental age.

Methods

Participants. A sample of 226 children were recruited at the Children's Discovery Museum of San Jose. 86 children were excluded because they did not satsify our planned inclusion criteria: within the age range of 24-48 months ($n = 13$), completed all trials ($n =$ 66), exposed to English greater than 75% of the time ($n = 37$), and correctly answered at least half of the familiar noun control trials ($n = 55$). Our final sample included 140 children ($N_{\text{females}} = 87$).

Stimuli. The disambiguation task included color pictures of 14 novel objects (e.g., a pair of tongs) and 24 familiar objects (e.g. a cookie; see SI). Items in the vocabulary assessment were a fixed set of 20 developmentally appropriate words from the Pearson Peabody Vocabulary Test (L. M. Dunn, Dunn, Bulheller, & Häcker, 1965). The novel words were XXX.

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. 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 iPad and asked to select one, given a label. If the participant chose incorrectly on a practice trial, the audio would correct them and allow the participant to choose again.

The child then completed the test phase. Like the practice trials, each of the test trials consisted of a word and two pictures, and the child's task was to identify the referent. 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, children were given a novel word and expected to select the novel object via the disambiguation inference. On familiar referent trials, children were given a familiar word and expected to select the correct familiar object. On Novel-Familiar trials, children saw a picture of a novel object and a familiar objects (e.g. a cookie and a pair of tongs). On Novel-Novel trials, children saw pictures of two novel objects (e.g. a pair of tongs and a leak) [How were N words introduced for NN trials?]. 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). 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 2 practice trials followed by 20 test trials.

Data analysis. 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 [0.88, 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 = 2.32 [2.02, 2.63]; Control-NN: M = 0.78, SD = 0.25, d = 1.1 [0.85, 1.35]). 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 = 1.62 [1.34, 1.89]; NN: M = 0.79, SD = 0.27, d = 1.08 [0.83, 1.33]).

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 = -0.89, SE = 0.26, Z = -3.41, p < .001). The main effect of target type was not significant (B = -0.47, SE = 0.3, Z = -1.6, p = 0.11). The interaction between the two factors was marginal (B = 0.72, SE = 0.38, Z = 1.9, p = 0.06), 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 assesessment. 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 = 0.45 [0.3, 0.57], $p < .001$).

term	Beta	SE	Z	p
(Intercept)	2.04	0.19	10.74	<.0001
Vocabulary	0.85	0.16	5.30	<.0001
Trial Type (NN)	-0.18	0.28	-0.65	0.51
Age	0.16	0.16	0.99	0.32
Vocabulary x Trial Type (NN)	-0.36	0.23	-1.57	0.12
Vocabulary x Age	0.01	0.14	0.09	0.93
Age x Trial Type (NN)	0.02	0.22	0.08	0.93
Vocabulary x Age x Trial Type (NN)	0.00	0.20	-0.02	0.98

Fig. **??** 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 1 presents the model parameters. The only reliable predictor of accuracy was vocabulary size (B = 0.85, SE = 0.16, Z = 5.3, 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 = 0.16, SE = 0.16, Z = 0.99, p = 0.32).

Discussion. Experiment 1 directly examines the relationship between the strength of the disambiguation effect and vocabulary size. We find that the strength of the 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 magnitude of the effects that we find are roughly consistent with meta-analytic estimates of those same effects. Figure X 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.[3] The change in effect size between As in the meta-analytic models, the effect

[3]Because some participants had no variability in their responses (all correct or all incorrect), we used the

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

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 test a causal relational between vocabulary knowlege and the disambiguation effect through an experimental manipulation.

Experiment 2: ME and Familiarity

In Experimtnet 2, we test a causal relationship between vocabulary size and the disambiguation effect by experimentally manipulating the strength of word knowledge. We

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.

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 certainity about the label for an object, with objects that have been labeled more frequently leading to high certainty about the label name. This object then serves as the "familiar" object in a novel-novel trial. If the strength of vocabulary knowledge about the "familiar" object influenced

Methods

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

Participants.

age_group	mean_age	n
2	30.98684	38
3	40.98571	35
4	52.16216	37

We planned a total sample of 108 children, 12 per between-subjects labeling condition, and 36 total in each one-year age gorup. Our final sample was 110 children, ages Inf – -Inf 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 (Marchena, Eigsti, Worek, Ono, & Snedeker, 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, experimenter error, or a child who recognized the target object, chose both objects, or did not make a choice. 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. The analysis we report here is consistent with that used in (**???**), though there are some slight numerical differences due to reclassification of exclusions.

err_type	n	pct
changed mind	2	0.0045455
exp err	2	0.0045455
interference	11	0.0250000
no choice	8	0.0181818
recog obj	4	0.0090909

Results and Discussion

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3076191	0.1046804	2.938650	0.0032965
age_mo_c	0.0464060	0.0112418	4.127972	0.0000366
times_labeled_c	0.4832010	0.1287155	3.754022	0.0001740
age_mo_c:times_labeled_c	0.0214303	0.0135810	1.577960	0.1145749

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. We centered both age and number of labels for interpretability of coefficients. Model results are shown in Table XYZ. 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.

General Discussion

References

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., Borovsky, A., & Fernald, A. (2012). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30months. *Cognition*.

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

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.

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.

Golinkoff, R., 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., 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.

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

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

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

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

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

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

Markman, E., 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.

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

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.

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.

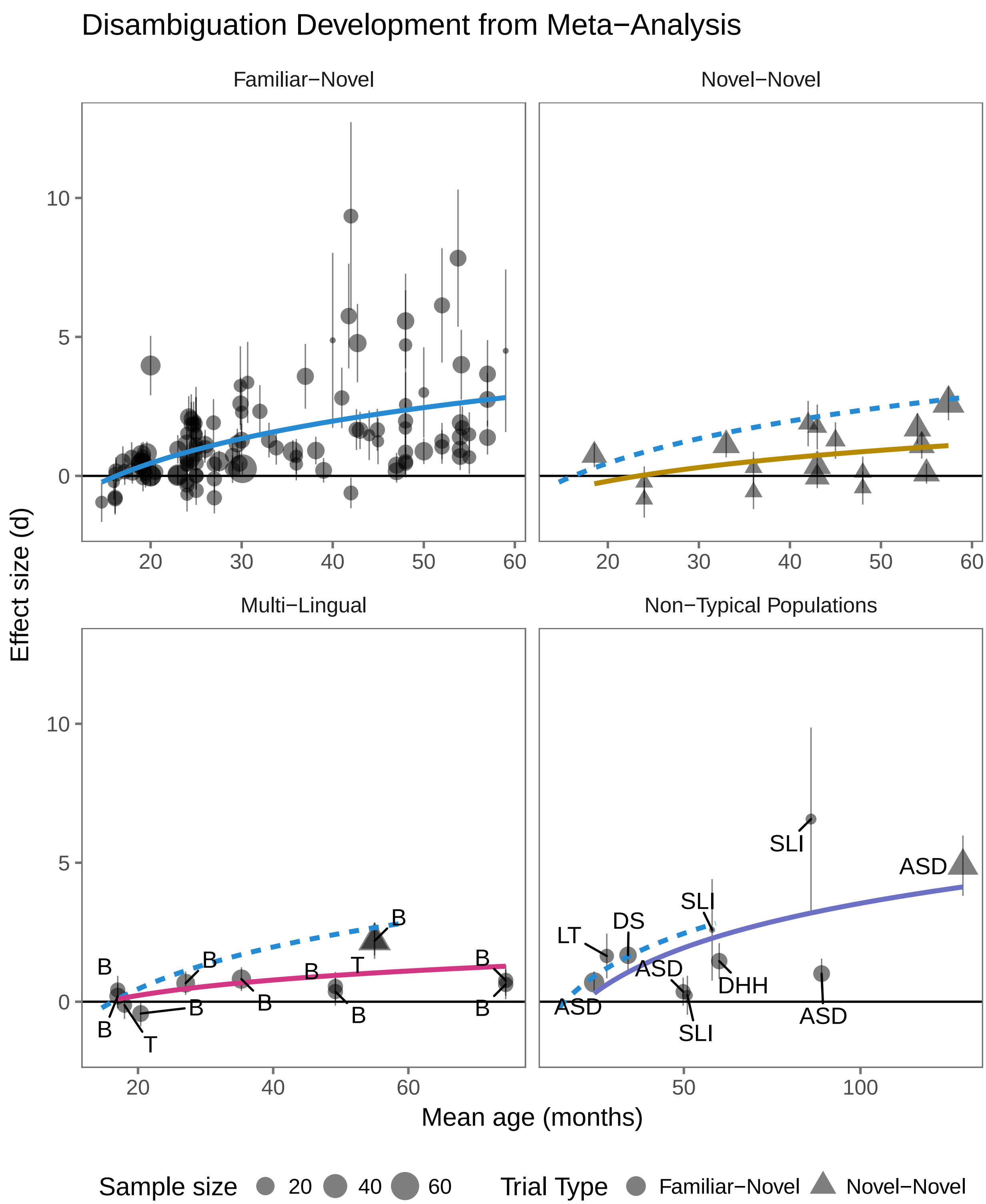


Figure 2. Developmental plots for each moderator. Ranges correspond to 95% confidence intervals. Model fits are log-linear. Point size corresponds to sample size, and point shape corresponds to trial type (NN vs. NF). Note that the x-axis scale varies by facet. B = bilingual; T = trilingual; LT = late-talker;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

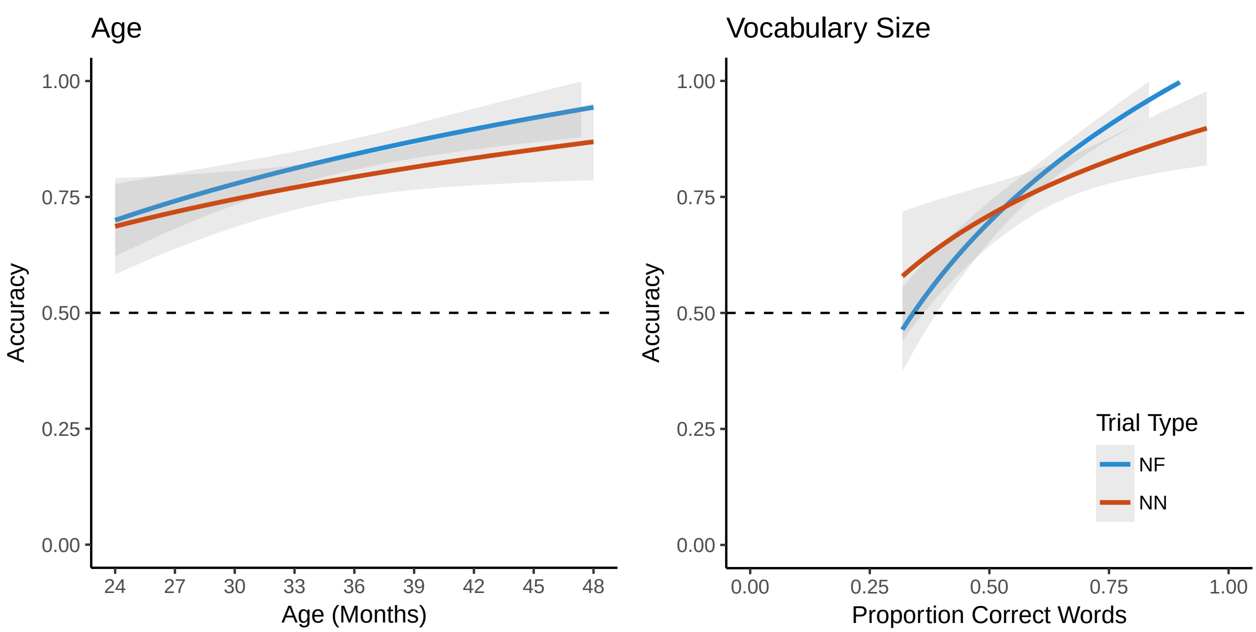


Figure 3. (#fig:test_p)Accuracy as a function of age (months; left) and vocabulary size (proportion correct on vocabulary assessment; right). Blue corresponds to trials with the canonical novel-familiar disambiguation paradigm, and red corresponds to trials with two novel alternatives, where a novel 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.

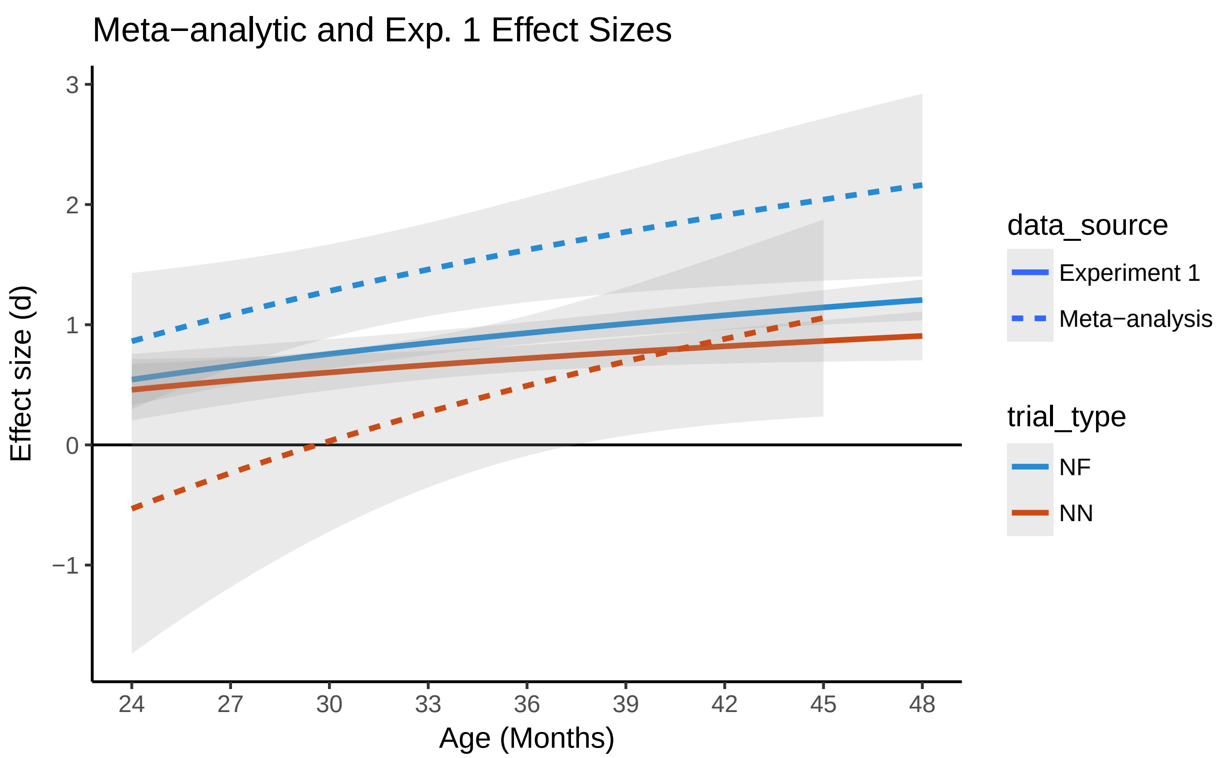
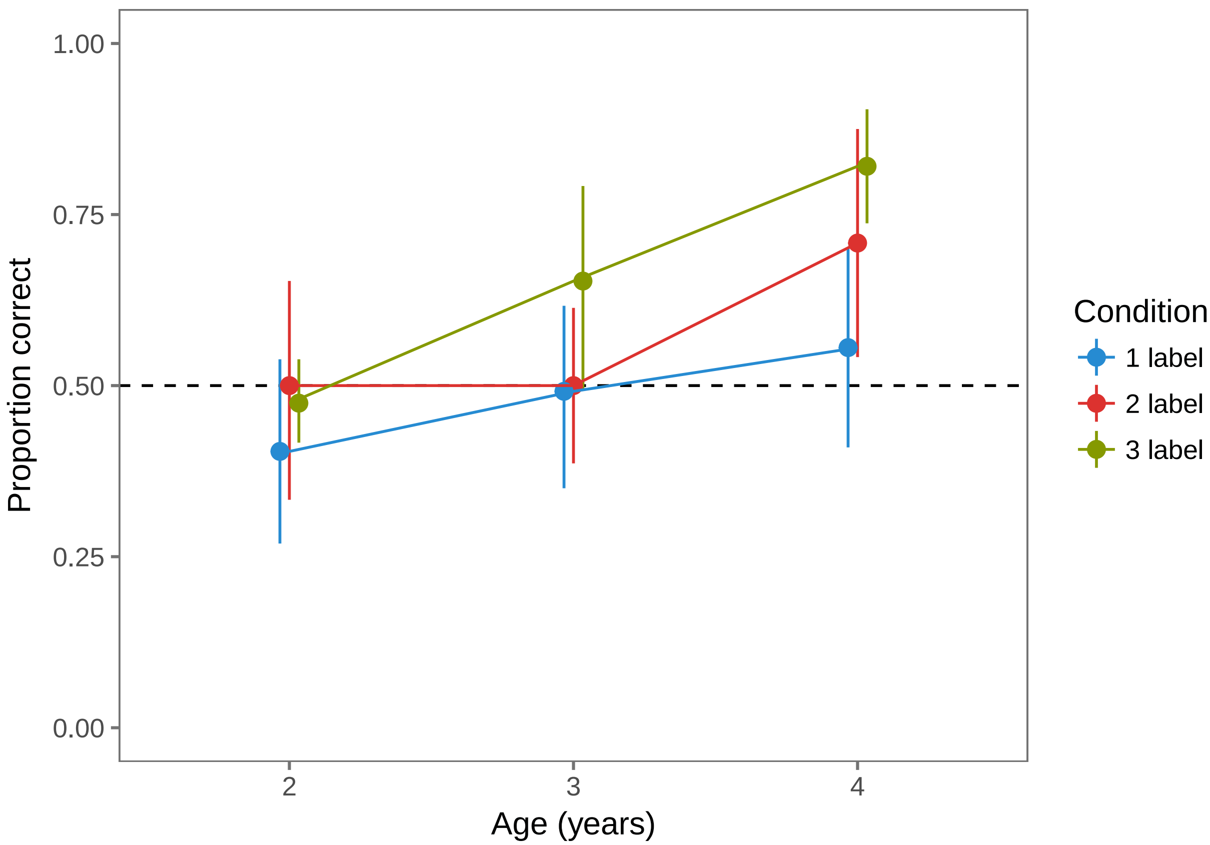


Figure 4. Meta-analytic data and data from experimental trials in Experiment 1 as a function of age. Effect sizes for Experiment 1 data are calculated for each participant, assuming the across-participant mean standard deviation as an estimate of the participant level standard deviation. Ranges are 95% confidence intervals.



1.00
0.75
0.50
0.25
0.00
Proportion correct
2
3
4
Age (years)
Condition
1 label
2 label
3 label

Figure 5