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. Theoretical accounts ofthis effect have debated whether it is due to initial constraints on children's lexicons (e.g. a principle of mutual exclusivity) or situation-specific pragmatic inferences. We present synthesis of exisiting evidence on this phenomonon through a meta-analysis of the existing literature. We then present two experiments that help distinguish between these theoretical constraints. We conclude by suggesting that multiple cognitive mechanisms may underlie the disambiguation effect in word learning.

Keywords: keywords

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; 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 (CHECK THESE CITES, including **???**, **???**, **???**), 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 (**???**). 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 and 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

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 that used one of two paradigms: (a) the canonical experimental paradigm for testing disambiguation behavior (an experimenter says a novel word in the context of a familiar object and a novel object, and the child guesses the intended referent; "Familiar-Novel"), or (b) a paradigm that exposed children to an unambigous mapping of a novel label to a novel object, and then introduced a second novel object and asked children 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 (MCDI; Fenson et al., 1994,Fenson et al. (2007)), referent type (object or picture), and number of alternatives in the forced choice task. We coded participant population as one of three subpopulatoins 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

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

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 coded several additional variables: 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 data were 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. Effect sizes were computed by a script, `compute_es.R`, available in the Github repository. We calculated effect sizes from reported means and standard deviations where available, otherwise we relied on reported test-statistics (t or d). All analyses were conducted with the metafor package (Viechtbauer, 2010) using mixed-effect models with grouping by paper. [2] In models with moderators, moderators variables were included as additive fixed effects.

Results

Bias. We first conducted analyses to determine if there was bias present in the literature. (see metalab)

Effect size estimates. To estimate the overall effect size, we fit a mixed effect model for the full sample of conditions. Effect size estimates are presented in Figure 1. The overall effect size estimate reliably differed from zero ($d = 1.06$ [0.81, 1.3]; see Appendix for by-condition forest plots). We also fit additional models for sub-populations in our sample ($d_{\text{typically-developing}} = 1.19$ [0.89, 1.49]; $d_{\text{multilingual}} = 1.19$ [0.89, 1.49]; $d_{\text{non-typically-developing}} =$

[2]The exact model specification was as follows: $model < -metafor :: rma.mv(yi = effect_{s}ize, V = effect_{s}ize_{v}ar, random = \ 1|paper, data = d)$.

1.57 [0.69, 2.44]). In the next set of analyses, we ask whether our moderator variables predict variance in these overall effect size estimates.

Trial type, referent type, and number of alternatives. Here we examined the influnce of three methodological varirables – trial type, referent type and number of alternatives – on effect size. Trial type was a signifcant predictor of effect size, with Familiar-Novel conditions leading to overall larger effect sizes, compared to Novel-Novel conditions (β = -0.71; Z = -4.65; p = <.001). Referent type (object vs. picture;) and number of alternatives () were not reliable predictors of effect size.

Age. We next examine developmental change in the magnitude of the disambiguation effect. See Table 1.

Vocabulary.

Bilingualism

Autism Spectrum Disorders

Experiment 1: ME and Vocabulary

Methods

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

Participants. Children were recruited at the Children's Discovery Museum of San Jose. Children were asked if they would be willing to play an iPad game with the experimenter and were informed that they could stop playing at any time. Children first completed two tasks adapted to iPad; one probing their vocabulary size and one mutual exclusivity inference task. Included in analyses are 166 children out of a planned sample of 160 participants. We ran 62 additional children, who were excluded from analysis based on planned exclusion criteria of low English language exposure (less than or equal to 75%), outside the age range of 24-48 month, children who do not give correct answers on > 50% of familiar noun (control) trials, or $< 100\%$ of trials completed. Included in our sample were 97 females and 69 males.

Stimuli. Mutual exclusivity inference task was comprised of 19 trials total; three practice trials of Familiar-Familiar (FF) nouns and 16 experimental trials. Experimental trials consisted of Novel-Familiar (NF), and Novel-Novel (NN) noun pairings. Of the pictures presented in the task, 14 objects were familiar and 24 objects were novel. The task included 8 control trials, equally split between NN noun pairings (C-NN) and NF noun pairings (C-NF) given in random order. Children who did not give correct answers on 50% of control trials were excluded from the final sample. The remaining 8 trials were divided equally between NN and NF trials.

The general format of the vocabulary assessment comprised of a 4 image display and a verbal prompt. Two practice trials were administered, followed by 20 experimental trials. Experimental trials included a fixed set of 20 developmentally appropriate words taken from the Pearson Peabody Vocabulary Test. These words were taken from 9 different domains, including professions, food, outside things, instruments, animals, classroom, shapes, verbs, and household items.

Procedure. Sessions took place individually in a small testing room away from the museum floor. In the ME inference task, the experimenter introduced them 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 saw 3 practice trials with two commonly known objects (i.e. cup and cookie). If the participant chose incorrectly for this practice trial, the audio would correct them and allow the participant to choose again. After the practice trials were completed, the task proceeded to run 16 test trials. Reaction times were measured from the onset of the target word. Children could only make one selection. The vocabulary task displayed 4 images randomly selected from the fixed bank of 22 images. Participants were prompted to choose one object. Again, reaction times were measured from the onset of the target word and children could only make one selection.

Data analysis. We used R (3.4.1, R Core Team, 2017) for all our analyses.

Results and Discussion

Could be specific strength of particular word in the NF pairing

but we also get it for NN trials alone

Experiment 2: ME and Familiarity

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 (**???**). 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 Lewis and Frank (2013), 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.

ME in Models of Word Learning

Basic statistical biases ("explaining away")

Regier (2005) model shows ME emergent

as noted by Frank, Goodman, Lai, and Tenenbaum (2009), Yu and Ballard (2007) model (IBM machine translation model #1, (**???**) for that; subsequently adapted by Nematzadeh, Fazly, and Stevenson (2012)) shows ME as well.

this is because any conditional probability model will show the same effect

In other words, Markman and Wachtel (1988)'s sense of a basic inductive bias will likely be present in a wide variety of different learning models.

What is the experience-dependence of ME in these models? In the Frank et al. (2009) model, the strength of the ME response scales with the strength of the familiar word's mapping; the same thing is true for the other models presumably.

Open question whether the actual differerence in a 2-year-olds' and a 4-year-olds' strength of representation of "ball" is what matters here?

Frank et al. (2009) model shows ME, in fact stronger than basic conditional probability. This is in part due to the use of the intention variable.

As a side note, the (**???**) no retention finding is shown in an even more pragmatic model: Smith, Goodman, and Frank (2013) model shows ME with no retention (though explanation in that model is a little implausible "because the speaker might not be committed to that label and is just using it as a matter of convenience.")

Primary point: No support here for overhypothesis building, which is suggested by 1) the bilingualism results. In order to fit the bilingual data, in general we'd have to assume that strength of individual representations in monolinguals and bilinguals was a driver, and this seems unlikely. 2) no support for E1 vocab findings unless the entire developmental trend is due to strength of the familiar word representations. In general, the strong — likely false — claim from all of these models is that the individual representation of the familiar object strength is the only locus for developmental/population-related change.

McMurray, Horst, and Samuelson (2012) model has ME emerge from the competition dynamics of a neural network.

> Thus, the selection of the novel object is dependent on the learning rule, but not because the network needs to learn something about that object/word. Rather, the weights between the known word/objects and the unused lexical units must decay, and the weights between the novel ones must not in order to create a platform upon which real-time competition dynamics can select the right object. A different type of weight decay (for example, if all weights decayed on each epoch) would not preserve the right form of the weight matrix. However, learning is not the whole story: this pattern of connectivity could not be harnessed in situation time without the gradual settling process represented by the inhibition and feedback dynamics. Moreover, the model's ability to learn from M.E. referent selection may also depend on this competition/feedback cycle. The model must select a single lexical unit and selectively amplify the novel object in order to

> eventually turn a word-referent link created during M.E. referent selection into a known word by associating the novel object with the novel word over many instances. Thus, while as a real-time process mutual exclusivity is likely to impact learning, it is really more the product of learning than a mechanism of it.

This proposal is complicated but might capture the global and local dynamics in Experiment 1 & 2 better than others.

(**???**) deal with bilingual data by adding a direct ME-related penalty, not letting it be emergent.

General Discussion

References

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., & 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.

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

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

Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In *Proceedings of the 31st annual meeting of the cognitive science society*.

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.

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

Markman, E. (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.

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.

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.

Nematzadeh, A., Fazly, A., & Stevenson, S. (2012). A computational model of memory, attention, and word learning. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics* (pp. 80–89). Association for Computational Linguistics.

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.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

Smith, N. J., Goodman, N., & Frank, M. (2013). Learning and using language via recursive pragmatic reasoning about other agents. In *Advances in neural information processing systems* (pp. 3039–3047).

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149–2165.

Model	Fixed effect	beta	z-value	p-value
Grand estimate	intrcpt	-0.14 [-0.47, 0.18]	-0.85	0.39
Grand estimate	ME_trial_typeNN	-0.92 [-1.21, -0.62]	-6.13	<.001
Grand estimate	mean_age	0.04 [0.03, 0.04]	11.96	<.001
Typical populations	intrcpt	-0.08 [-0.43, 0.28]	-0.42	0.68
Typical populations	ME_trial_typeNN	-1.08 [-1.38, -0.78]	-7.05	<.001
Typical populations	mean_age	0.04 [0.04, 0.05]	12.34	<.001
Multilingual populations	intrcpt	-0.07 [-0.62, 0.48]	-0.26	0.79
Multilingual populations	ME_trial_typeNN	1.64 [0.72, 2.56]	3.49	<.001
Multilingual populations	mean_age	0.01 [0, 0.03]	1.57	0.12
Non-typically developing populations	intrcpt	0.66 [-0.57, 1.89]	1.05	0.29
Non-typically developing populations	ME_trial_typeNN	3.06 [0.58, 5.55]	2.42	0.02
Non-typically developing populations	mean_age	0.01 [-0.01, 0.03]	0.77	0.44

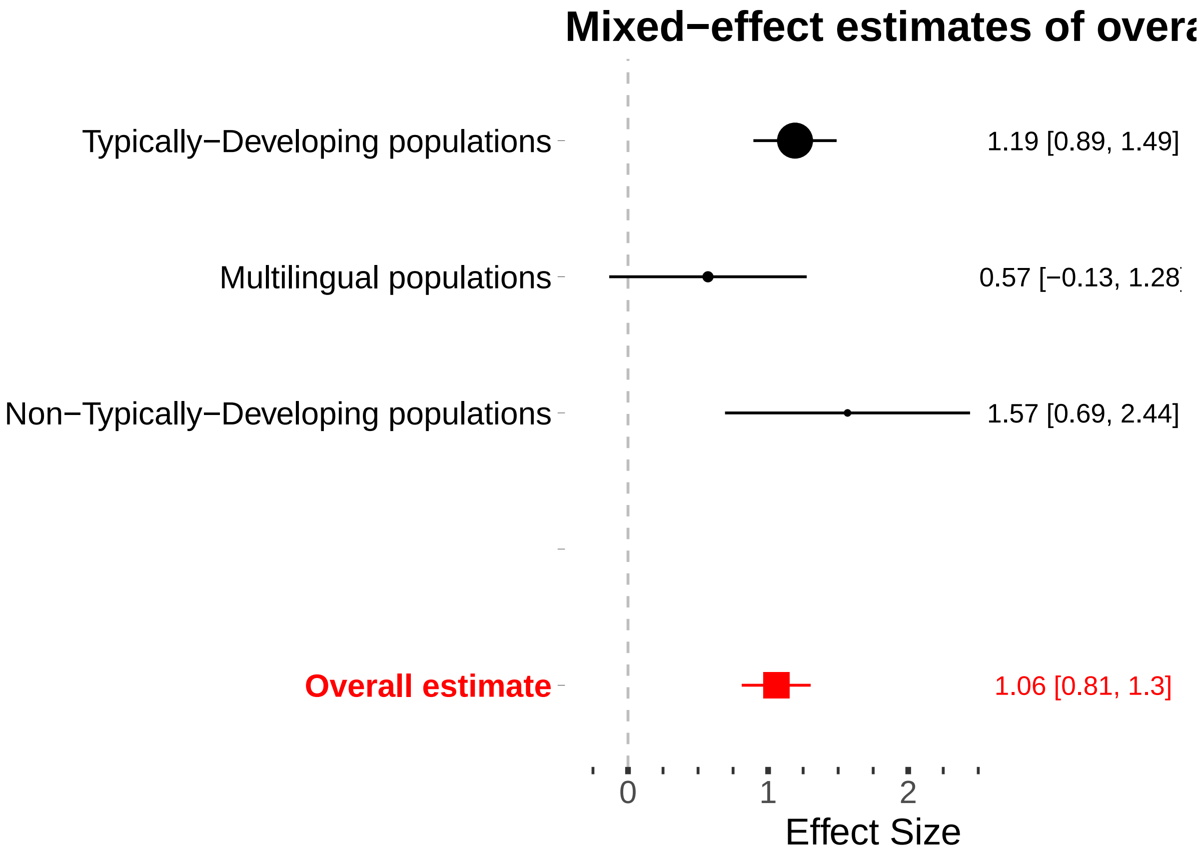


Figure 1. Mixed-effect effect size estimates for all conditions (red) and each of the three sub-populations in our sample. Ranges are 95% confidence intervals. The size of the point for the sub-populations corresponds to sample size.

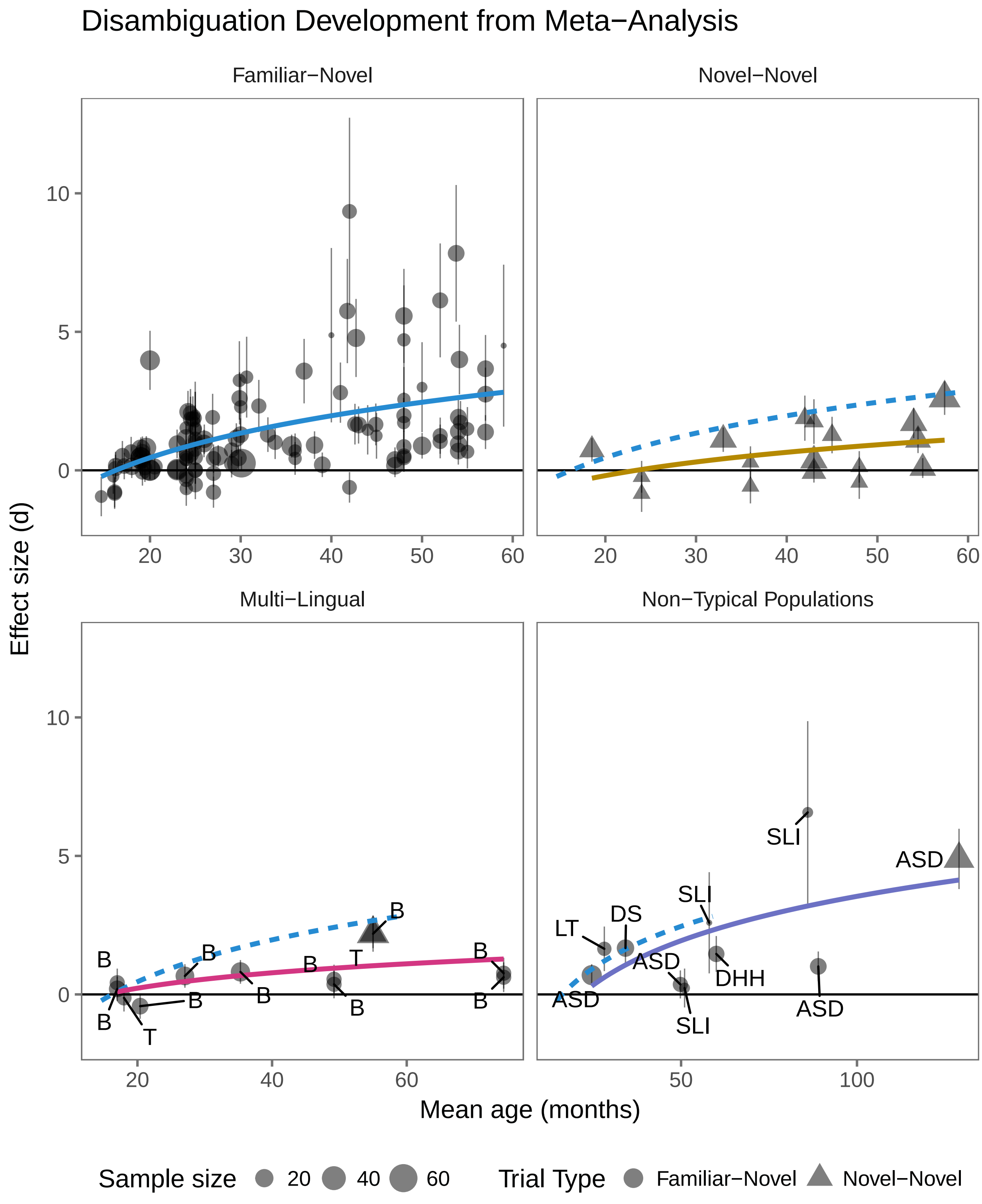
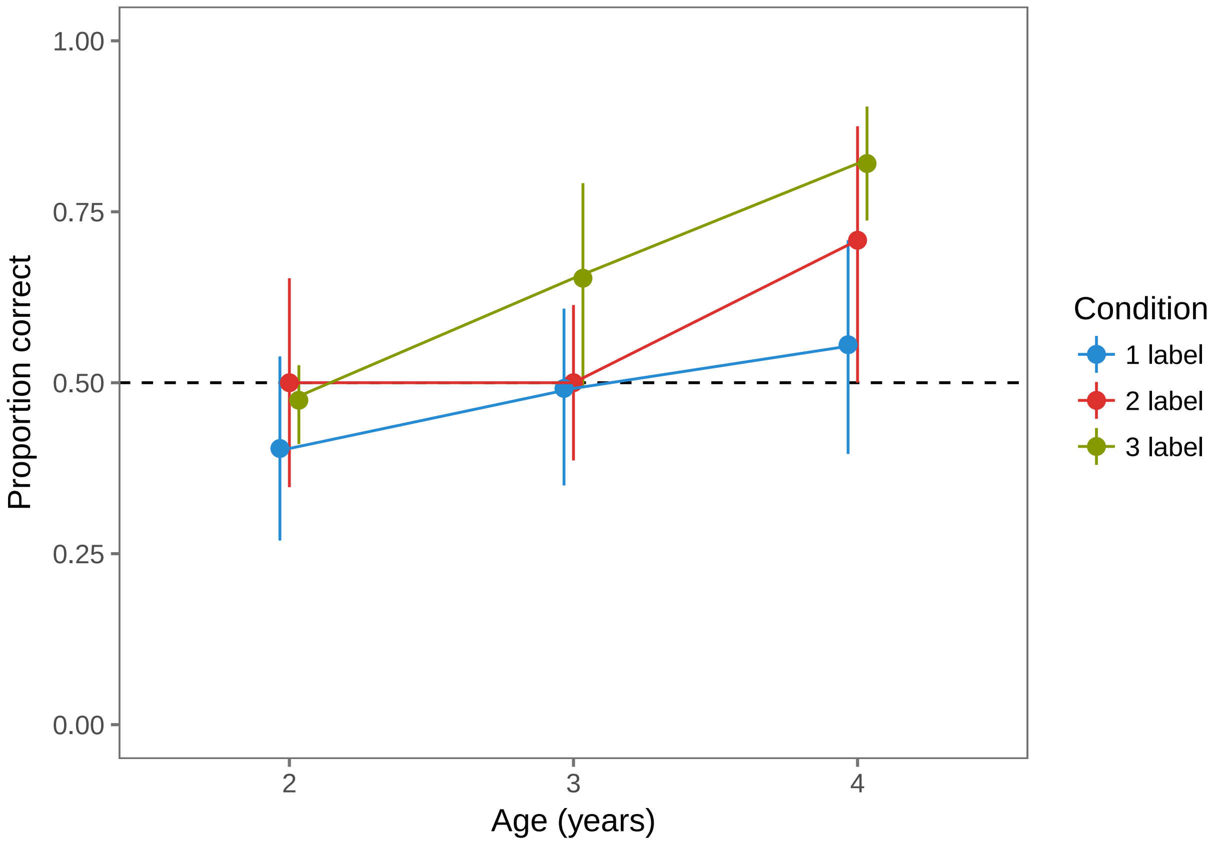


Figure 2. Developmental plots for each moderator. Ranges correspond to 95% confidence intervals. Model fits are log-linear.



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 3