A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Molly Lewis[1], Mika Braginsky[1], Sho Tsuji[2], Christina Bergmann[2], Page Piccinini[2], Alejandrina Cristia[2], Michael C. Frank[1]

[1] Department Psychology, Stanford University

[2] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS

Author note

Correspondence concerning this article should be addressed to Molly Lewis, Psychology Department, Stanford University. 450 Serra Mall, Stanford, CA 94305. E-mail: mll@stanford.edu.

Abstract

replicability, etc.

Keywords: replicability, reproducibility, meta-analysis, developmental psychology, language acquisition

Word count: XXXX

A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Introduction

To learn to speak a language, a child must acquire a wide range knowledge and skills: the sounds of the language, the word forms, and how words map to meanings, to name only a few. How does this process unfold? Our goal as psychologists is to build a theory that can explain and predict this process in a way that is both precise but also highly generalizable. The challenge we face, however, is that we must build this theory on the basis of very limited data, and, as a consequence, we must rely on error-prone inductive reasoning strategies to build our theories. In this paper, we consider the domain of language acquisition and demonstrate how meta-analytic methods can support the inductive theory building process.

Meta-analysis is a quantitative method for aggregating across experimental findings. The fundamental unit of meta-analysis is the *effect size*: a scale-free, quantitative measure of "success" in a phenomenon. Importantly, an effect size provides an estimate of the *size* of an effect, as well as a measure of uncertainty around this point estimate. With such a quantitative measure of success, we can apply the same reasoning we use to aggregate noisy measurements over participants in a single study: By assuming each *study*, rather than participant, is sampled from a population, we can appeal to the classical statistical framework to combine estimates of the effect size for a given phenomenon.

Meta-analytic methods support theory building in several ways. First, they provide a way of evaluating which effects in a literature are "real," and thus should constrain the theory. This is particulary important in light of recent high-profile evidence in the field that an effect observed in one study may not replicate in another (Collaboration & others, 2012, 2015; "replication crisis", Ioannidis, 2005). Failed replications are difficult to interpret, however, because they may result from a wide variety of causes, including an in initial false positive, a subsequent false negative, or differences between initial and replication studies, and making causal attributions in a situation with two conflicting studies is often difficult (Gilbert et al., 2016; Anderson et al., 2016). Meta-analysis can allow researchers to address

this set of issues in a principled way by aggregating evidence across studies and assuming that there is some variability in true effect size from study to study. In this way, meta-analytic methods can provide more veridical description of the empirical landscape, which in turn leads to better theory-building.

Second, meta-analysis supports theory building by providing higher fidelity descriptions of phenomenona. Rather than simply concluding that an effect exists, effect sizes allow us to ask finer grain questions: How much variability is there in the effect? How does the effect change over development? To what extent does a moderator of theoretetic influence the effect? This type of continuous analysis supports building quantitative models, and specifying theories that are more precise and constraining.

Furthermore, effect sizes provide a common language of comparing *across* phenomena. This common language allows us to meaningfully consider the relationship between different phenomena in the language acquisition domain ("meta-meta-analysis"). Through cross-phenomena comparions, we can understand not only the trajectory of a particular phenomenon, like word learning for example, but also how this phenomenon might depend on other skills, such as sound learning, gaze following, and many others. With this more complete picture of the interaction of different linguistic competencies, we can begin to build a more synthetic theory of language acquisition.

Finally, in addition to these theoretical motivations, there are practical reasons for conducting a quantitative synthesis. When planning an experiment, an estimate of the size of an effect on the basis of prior literature can inform the sample size needed to achieve a desired level of power. Meta-analytic estimates of effect sizes can also aid in design choices: If a certain paradigm tends to have overall larger effect sizes than another, the strategic researcher might select this paradigm in order to maximize the power of a study.

While meta-analytic methods are likely helpful for many psychological literatures, we believe language acquisition is a particularly informative application for this tool. One reason is that language acquistion may be uniquely vulnerable to false findings because

running children is expensive, and thus sample sizes are small and studies are underpowered (Ioannidis, 2005). In addition, the difficulty in running participants means that replications are relatively rare in the field. Finally, there has been attention to developmental psychology research practices more broadly, suggesting evidence of experimenter bias (Peterson, 2016).

We take as our ultimate goal a single overarching theory of language acquisition that can explain and predict all the relevant pheonemona. Toward this end, we developed a dataset of effect sizes in the language acquistion literature across 12 different phenomenona (Metalab; http://metalab.stanford.edu/). We demonstrate how meta-analysis supports building this theory in two ways. We first use meta-analytic techniques to evaluate the evidential value of the empirical landscape in language acquistion research. We find broadly that this literature has strong evidential value, and thus that the effects report in the literature should constrain our theory of language acquisition. We then turn to synthesizing these findings across phenenomena and offer a preliminary theoretical synthesis of the field.

Method

We analyzed 12 different phenomenena in language acquisition. These phenomena were selected opportunistically, based on their availability in the literature and feasibility of conduciting meta-analysis for a particular phenomenon. The phenomenena cover development at many different levels of the language hierarchy, from the acquistion of prosody and phonemic contrasts, to gaze following in linguistic interaction. This wide range of phenomena allowed us to compare the course of development across different domains, as well as explore questions about the interactive nature of language acquisition (Table 1).

To obtain estimates of effect size, we coded papers reporting experimental data. Within each paper, we calculated a separate effect size estimate for each experiment and age group ("condition"). In total, our sample includes estimates from 269 papers, 981 different conditions and 12,029 participants. The process for selecting papers from the literature differed by domain, with some individual meta-analyses using more systematic approaches

than others (see SI).

Level	Phenomenon	Description	N papers (conditions)
Prosody	IDS preference (Dunst, Gorman, & Hamby, 2012)	Looking times as a function of whether infant-directed vs. adult-directed speech is presented as stimulation.	16 (50)
Sounds	Phonotactic learning (Cristia, in prep.)	Infants' ability to learn phonotactic generalizations from a short exposure.	15 (47)
	Vowel discrimination (native) (Tsuji & Cristia, 2014)	Discrimination of native-language vowels, including results from a variety of methods.	40 (167)
	Vowel discrimination (non-native) (Tsuji & Cristia, 2014)	Discrimination of non-native vowels, including results from a variety of methods.	21 (72)
	Statistical sound learning (Cristia, in prep.)	Infants' ability to learn sound categories from their acoustic distribution.	11 (40)
	Word segmentation (Bergmann & Cristia, 2015)	Recognition of familiarized words from running, natural speech using behavioral methods.	68 (296)
Words	Mutual exclusivity (Lewis & Frank, in prep.)	Mapping of novel words reflecting children's inference that novel words tend to refer to novel objects.	20 (60)
	Sound Symbolism (Lammertink et al., in prep.)	Non-arbitrary relationship between form and meaning ("bouba-kiki effect").	10 (42)
	Concept-label advantage (Lewis & Long, unpublished)	Infants' categorization judgments in the presence and absence of labels.	16 (100)
	Online word recognition (Frank, Lewis, & MacDonald, 2016)	Online word recognition of familiar words using two-alternative forced choice preferential looking.	12 (32)
Communication	Gaze following (Frank, Lewis, & MacDonald, 2016)	Gaze following using standard multi-alternative forced-choice paradigms.	15 (45)
	Pointing and vocabulary (Colonnesi et al., 2010)	Longitudinal correlations between declarative pointing and later vocabulary.	25 (30)

Table 1
Overview of meta-analyses in dataset.

Replicability of the field

To assess the replicability of language acquisition phenomena, we conducted several diagnostic analyses: Meta-analytic estimates of effect size, fail-safe-N (Orwin, 1983), funnel plots, and p-curve (U. Simonsohn, Nelson, & Simmons, 2014; Simonsohn, Nelson, & Simmons, 2014; Simonsohn, Simmons, & Nelson, 2015). These analytical approaches each have limitations, but taken together, they provide converging evidence about the replicability of a literature. Overall, we find little evidence of bias in our meta-analyses, suggesting that the language acqusition literature likely describes real psychological phenomenona and should therefore provide the basis for theoretical development.

Meta-Analytic Effect Size

To estimate the overall effect size of a literature, effect sizes are pooled across papers to obtain a single meta-analytic estimate. This meta-analytic effect-size can be thought of as the “best estimate" of the effect size for a phenomenon given all the available data in the literature.

Table 2, column 2 presents meta-analytic effect size estimates for each of our phenomenona. We find evidence for a non-zero effect size in 11 out of 12 of our phenomena, suggesting these literature provide evidential value. In the case of phonotatic learning, however, we find that the meta-analytic effect size estimate does not differ from zero, suggest that this literature does not describe a robust effect. [Remove it from analyses below?].

While the meaure of effect size is itself quantitative, meta-analytic estimates of effect size provide only categorical information about the evidential value of a literature: the effect is real, or not. But, a more powerful method of assessing evidential value would tell us the *degree* to which a literature has evidential value, and thus the degree to which it should constrain our theory building. In the following three analyses—fail-safe-N, funnel plots, and p-curves—we describe through analyses that quantify the evidential value of these literatures.

Fail-safe-N

One approach for quantifying the reliability of a literature is to ask, How many missing studies with null effects would have to exist in the "file drawer" in order for the overall effect size to be zero? This is called the "fail-safe" number of studies (Orwin, 1983). To answer this question, we estimated the overall effect size for each phenomenenon (Table 2, column 2), and then used this to estimate the fail-safe-N (Table 2, column 3).

Because of the large number of positive studies in many of the MAs we assessed, this analysis suggests a very large number of studies would have to be "missing" in each literature ($M = 3634$) in order for the overall effect sizes to be 0. Thus, while it is possible that some reporting bias is present in the literature, the large fail-safe-N suggests that the literature nonetheless likely describes a real effect.

One limitation of this analysis, however, is that it assumes that all reported effect sizes are obtained in the absence of analytical flexibility: If experimenters are exercising analytical flexibility through practices like p-hacking, then the number and magnitude of observed true effects in the literature may be inflated. In the next analysis, we examine this possibility through funnel plots.

Funnel Plots

Funnel plots provide a visual method for evaluating whether variability in effect sizes is due only to differences in sample size. A funnel plot shows effect sizes versus a metric of sample size, standard error. If there is no bias in a literature, we should expect studies to be randomly sampled around the mean, with more variability for less precise studies.

Figure 1 presents funnel plots for each of our 12 meta-analyses. These plots show evidence of asymmetry (bias) for several of our phenomenon (Table 2, column 4). However, an important limitation of this method is that it is difficult to determine the source of this bias. One possibility is that this bias reflects true heterogenity in phenomena (e.g. different ages). P-curve analyses provide one method for addressing this issue, which we turn to next.

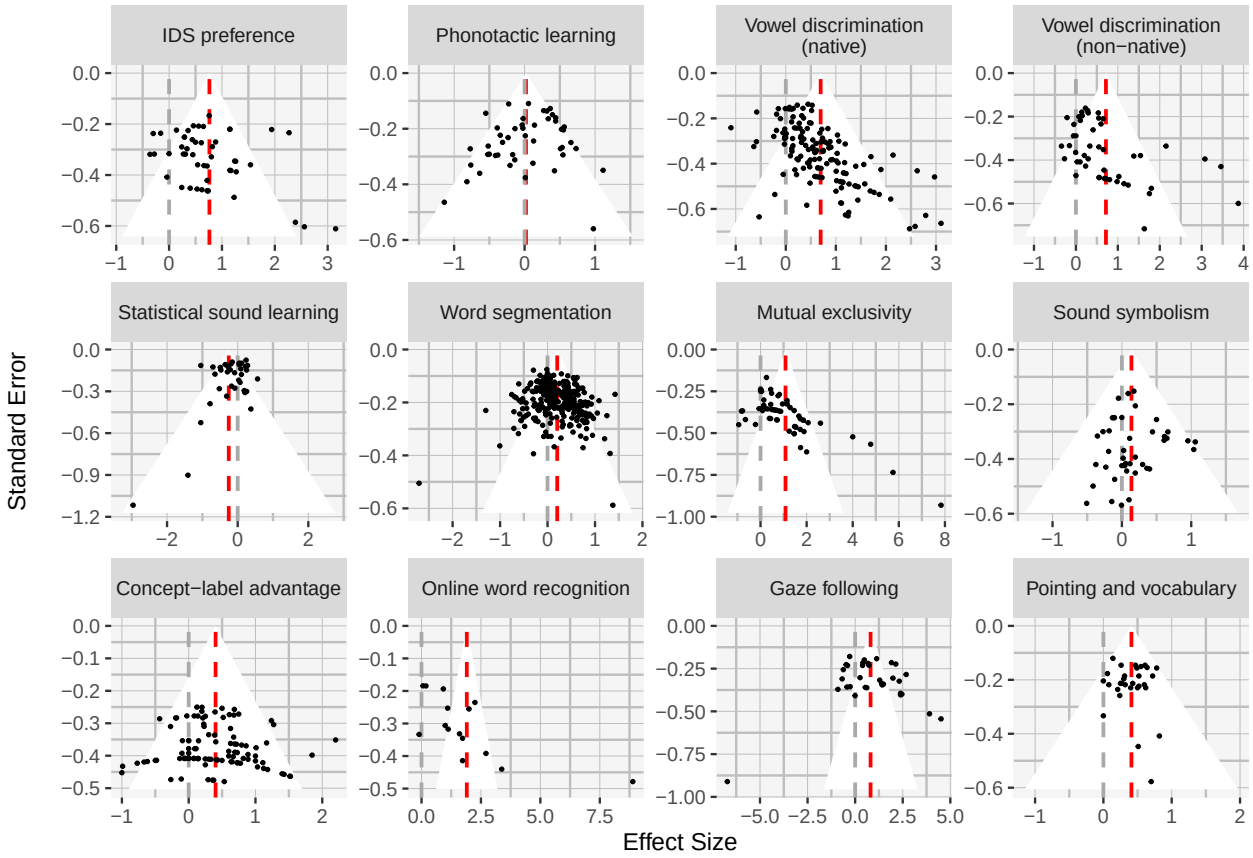

Figure 1. Funnel plots for each meta-analysis. Each effect size estimate is represented by a point, and the mean effect size is shown as a red dashed line. The funnel corresponds to a 95% (narrow) and 99% (wide) CI around this mean. In the absense of true heterogenity in effect sizes (no moderators) and bias, we should expect all points to fall inside the funnel.

P-curves

A p-curve is the distribution of p-values for the statistical test of the main hypothesis across a literature (U. Simonsohn et al., 2014; Simonsohn et al., 2014, 2015). Critically, if there is a robust effect in the literature, the shape of the p-curve should reflect this. In particular, we should expect the p-curve to be right skewed with more small values (e.g., .01) than large values (e.g., .04). An important property of this analysis is that we should expect this skew independent of any true heterogenity in the data, such as age. Evidence that the curve is in fact right-skewed would suggest that the literature is not biased, and that it provides evidential value for theory building.

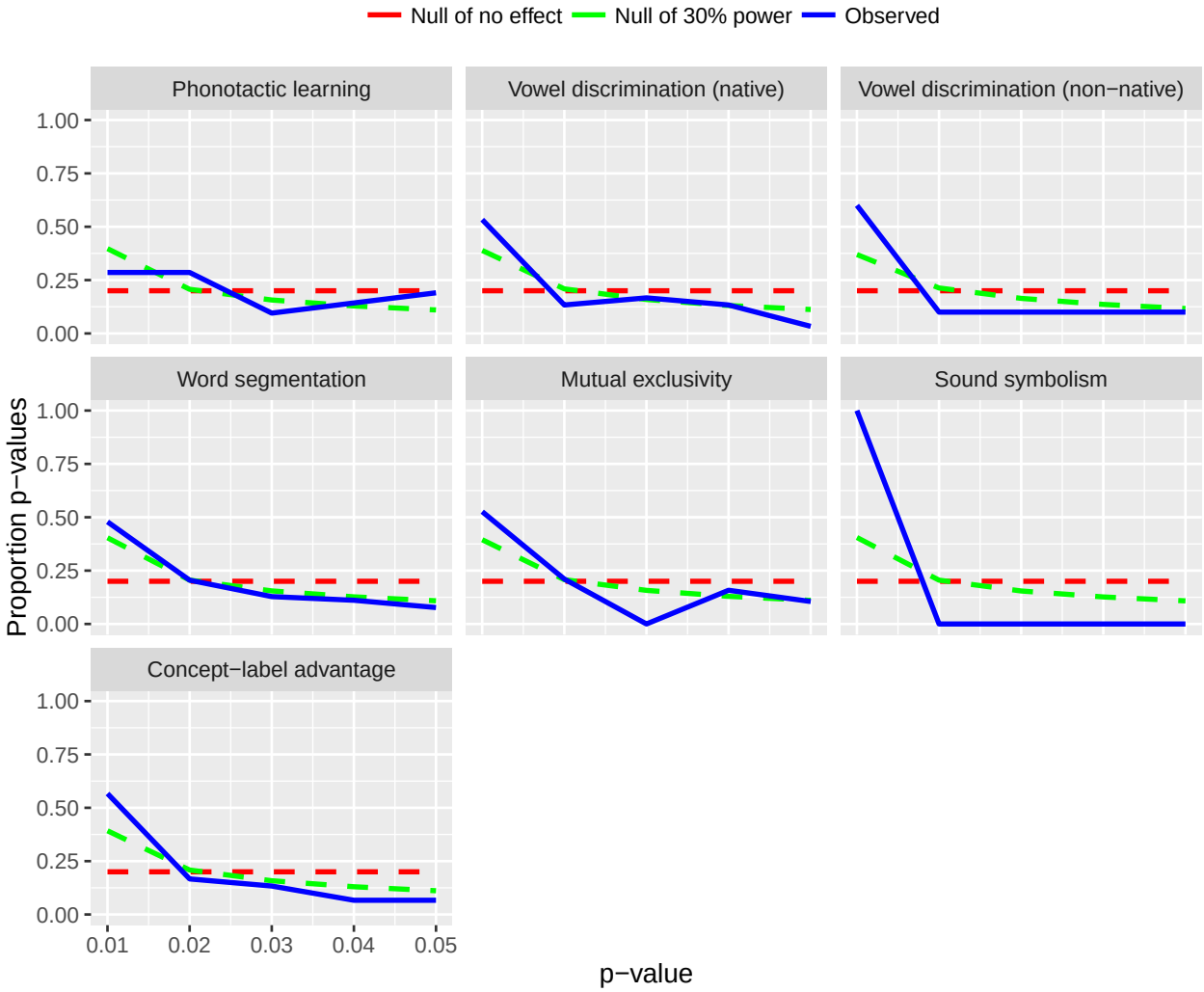

Figure 2. P-curve for each meta-analysis (Simonsohn, Nelson, & Simmons, 2014). In the absense of p-hacking, we should expect the observed p-curve (blue) to be right-skewed (more small values). The red dashed line shows the expected distribution of p-values when the effect is non-existent (the null is true). The green dashed line shows the expected distribution if the effect is real, but studies only have 33% power.

Phenomenon	*d*	fail-safe-N	funnel skew	p-curve skew	power
IDS preference	0.71 [0.53, 0.89]	3762	1.88 (0.06)		
Phonotactic learning	0.04 [-0.09, 0.16]	45	-1.08 (0.28)	-1.52 (0.06)	0.14
Vowel discrimination (native)	0.6 [0.5, 0.71]	9536	8.98 (0)	-5.42 (0)	0.67
Vowel discrimination (non-native)	0.66 [0.42, 0.9]	3391	4.13 (0)	-3.24 (0)	0.78
Statistical sound learning	-0.14 [-0.27, -0.02]	Inf	-1.87 (0.06)		
Word segmentation	0.2 [0.15, 0.25]	5645	1.54 (0.12)	-9.67 (0)	0.56
Mutual exclusivity	1.01 [0.68, 1.33]	6443	6.25 (0)		
Sound symbolism	0.15 [0.04, 0.26]	538	-1.32 (0.19)	-2.16 (0.02)	0.96
Concept-label advantage	0.4 [0.29, 0.51]	3928	0.31 (0.76)	-6.15 (0)	0.69
Online word recognition	1.89 [0.81, 2.96]	2843	2.92 (0)		
Gaze following	0.84 [0.26, 1.42]	2641	-1.69 (0.09)		
Pointing and vocabulary	0.41 [0.32, 0.49]	1202	0.59 (0.55)		

Table 2

Summary of replicability analyses. d = Effect size (Cohen's d) estimated from a random-effect model; fail-safe-N = number of missing studies that would have to exist in order for the overall effect size to be d = 0; funnel skew = test of asymmetry in funnel plot using the random-effect Egger's test (Stern & Eggers, 2005); p-curve skew = test of the right skew of the p-curve using the Stouffer method (Simonsohn, Simmons, & Nelson, 2015); power = power to reject the null hypothesis at the 5% significance level based on the p-curve (Simonsohn, Nelson, & Simmons, 2014); Brackets give 95% confidence intervals, and parentheses show p-values.

Figure 2 shows p-curves for 7 of our 12 meta-analyses.[1] With the exception of phonotactic learning, all p-curves show evidence of right skew. This is confirmed by formal analyses (Table 2, column 5).

P-curves also provide a method for calculcating the overall power of a literature, based on the shape of the p-curve (U. Simonsohn et al., 2014). Intuitively, when power is high and effect is real, we should be more likely to observe an effect size "further" from the null. This means that we will observe more small effect sizes. Thus, the higher the power, the more right skewed the p-curve will be. Table 2 (column 6) presents estimates of power for each

[1] We did not conduct p-curves on all meta-analyses because previously published meta-analyses did not include test statistics and the key test statistics in some others were inappropriate for p-curve.

meta-analysis based on p-curve. With the exception of phonotactic learning (*power* = .14), literatures appear to have acceptable power.

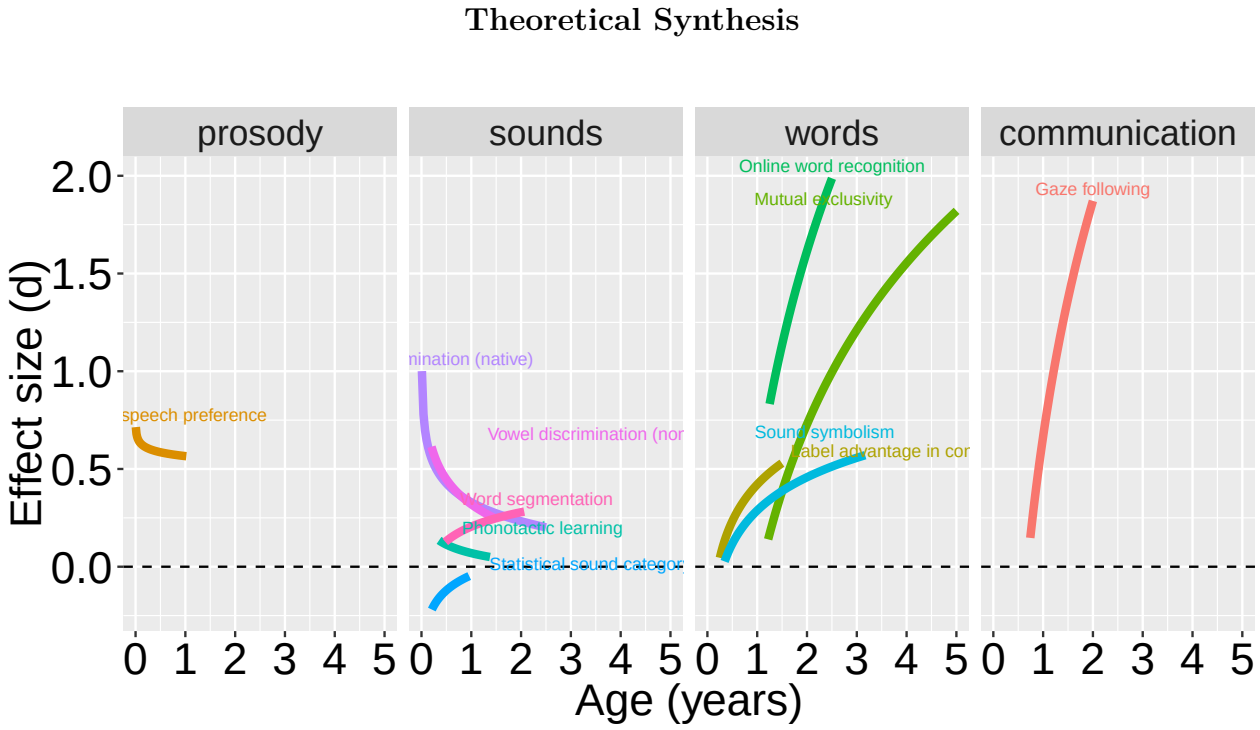

Figure 3. Meta-meta plot

- Ultimately, what we would like to have, is a theory of how children acquire language at all levels of the linguistic hierarchy.
- There are a number of different theories you could have.
- One possibilty is the "stages" hypothesis, where children learn language representations in a sequential fashion starting at the lowest levels of the hierarchy.
- Another possibility though is a more continous, synegistic theory, where learning at all levels is happening simultaneously.
- And there are a number of recent pieces of evidence that are consisntent with this. For example,
  - Infants learn phonetic contrasts when supported by word context (Feldman, et al., 2013).
  - Infants learn word mappings when supported by prosody (Shukla, White, & Aslin, 2011).

- But, the hypothesis space is really quite large, once you start thinking about it in terms of ES [plots of hypothesis space; e.g., https://speakerdeck.com/mllewis/metalab-1?slide=32]
  - For example, you might you learn low levels stuff like prosody first, then sounds, then words, then communicaiton
  - Another possibility is that you don't have to quite master a skill before learning gets off the ground
  - Learning also doesn't have to monotonic in these tasks – children in a task might get "worse," for example because the skill is no longer relevant (like non-native vowels)
- Fig. 3 shows what this looks like based our dataset
- Evidence for interactivity!

Discussion

Limitations

Author Contributions.

Acknowledgments.

References. Bergmann, C., & Cristia, A. (2015). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*.

Collaboration, O. S., & others. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660.

Collaboration, O. S., & others. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1).

Frank, M. C., Lewis, M. L., & MacDonald, K. (in press). A performance model for early word learning. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Retrieved from http://langcog.stanford.edu/papers_new/frank-2016-underrev.pdf

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124.

Lewis, M., & Frank, M. C. (in prep). Multiple routes to disambiguation.

Orwin, R. G. (1983). A fail-safe n for effect size in meta-analysis. *Journal of Educational Statistics*, 157–159.

Peterson, D. (2016). The Baby Factory: Difficult Research Objects, Disciplinary Standards, and the Production of Statistical Significance. *Socius: Sociological Research for a Dynamic World*, *2*(0), 1–10.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*,

9(6), 666–681.

Simonsohn, Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Simonsohn, Simmons, J. P., & Nelson, L. D. (2015). Better p-curves. *Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson (Forthcoming),"Better P-Curves," Journal of Experimental Psychology: General*.

Sterne, J. A., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments*, 99–110.

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191.