A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Molly Lewis*[1], Mika Braginsky[2], Sho Tsuji[3], Christina Bergmann[3], Page Piccinini[4], Alejandrina Cristia[3], Michael C. Frank[5]

[1] Computation Institute, University of Chicago

[2] Department of Brain and Cognitive Sciences, MIT

[3] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS

[4] NeuroPsychologie Interventionnelle, ENS

[5] Department of Psychology, Stanford University

Author note

*To whom correspondence should be addressed. E-mail: mollylewis@uchicago.edu

Abstract

To acquire a language, children must learn a range of skills, from the sounds of their language to the meanings of words. These skills are typically studied in isolation in separate research programs, but a growing body of evidence points to interdependencies across skills in the acquisition process. Here, we suggest that the meta-analytic method can support the process of building systems-level theories, as well as provide a tool for detecting bias in a literature. We present meta-analyses of 12 phenomena in language acquisition with over 700 effect sizes. We find that the language acquisition literature overall has a high degree of evidential value. We then present a quantitative synthesis of language acquisition phenomena that suggests interactivity across the system.

Keywords: developmental psychology, language acquisition, quantitative theories, meta-analysis

Word count: 4784

A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Introduction

Children beginning to acquire a language must learn its sounds, its word forms, and their meanings, and a number of other component skills of language understanding and use. A synthetic theory that explains the inputs, mechanisms, and timeline of this process is an aspirational goal for the field of early language learning. One important aspect of such a theory is an account of how the acquisition of individual skills depends on others. For example, to what extent must the sounds of a language be mastered prior to learning word meanings? Although a huge body of research addresses individual aspects of early language learning (see e.g., Kuhl, 2004 for review), only a small amount of work addresses the question of relationships between different skills (e.g., Feldman, Myers, White, Griffiths, & Morgan, 2013; Johnson, Demuth, Jones, & Black, 2010; Shukla, White, & Aslin, 2011). Yet if such relationships exist, they should play a central role in our theories.

The effort to build synthetic theories is further complicated by the fact that there is often uncertainty about the developmental trajectory of individual skills. Developmental trajectories are typically communicated via verbal (often binary) summaries of a set of variable experimental findings (e.g., "by eight months, infants can segment words from fluent speech"). In the case of contradictory findings then, theorists may be uncertain about which experimental findings can be used to constrain the theory, and often must resort to verbal discounting of one finding or the other based on methodological or theoretical factors. Resolving this issue requires a method for synthesizing findings in a more systematic and principled fashion.

We suggest that a solution to both of these challenges—building integrative whole-system views and evaluating evidential strength in a field of scientific research—is to describe experimental findings in quantitative, rather than qualitative, terms. Quantitative descriptions allow for the use of quantitative methods for aggregating experimental findings in order to evaluate evidential strength. In addition, describing experimental findings as

quantitative estimates provides a common language for comparing across phenomena, and a way to make more precise predictions. In this paper, we consider the domain of language acquisition and demonstrate how the quantitative tools of meta-analysis can support theory building in psychological research.

Meta-analysis is a quantitative method for aggregating across experimental findings (Glass, 1976; Hedges & Olkin, 2014). 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 this quantitative measure, 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 a statistical framework to combine estimates of the effect size for a given phenomenon.

Meta-analytic methods can support theory building in several ways. First, they provide a way to evaluate which effects in a literature are most likely to be observed consistently, and thus should constrain the theory. This issue is particularly important in light of recent evidence that an effect observed in one study may be unlikely to replicate in another (Ebersole et al., 2016; Open Science Collaboration, 2012, 2015). Failed replications are difficult to interpret, however, because they may result from a wide variety of causes, including an initial false positive, a subsequent false negative, or differences between initial and replication studies, such that making causal attributions in a situation with two conflicting studies is often difficult (Anderson et al., 2016; Gilbert, King, Pettigrew, & Wilson, 2016). By aggregating evidence across studies and assuming that there is some variability in true effect size from study to study, meta-analytic methods can provide a 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 phenomena. Given an effect size estimate, meta-analytic methods provide a method for quantifying the amount variability around this point estimate. Furthermore, the quantitative

framework allows researchers to measure potential moderators in effect size. This ability is particularly important for developmental phenomena because building a theory requires a precise description of changes in effect size across development. Individual papers typically describe an effect size for 1-2 age groups, but the ultimate goal for the theorist is to detect a moderator—age—in this effect. Given that moderators always require more power to detect (Button et al., 2013), it may be quite difficult to identify size from individual papers. By aggregating across papers using meta-analytic methods, however, we may be better able to detect these changes, leading to more precise description of the empirical phenomena.

Finally, effect size estimates also provide a common language for comparing across phenomena. In the current work, this common language allows us to consider the relationship between different phenomena in the language acquisition domain ("meta-meta-analysis"). Through cross-phenomenon comparisons, we can understand not only the trajectory of a particular phenomenon, such as word learning, but also how the trajectory of each phenomenon might relate to other skills, such as sound learning, gaze following, and many others. This more holistic description of the empirical landscape can inform theories about the extent to which there is interdependence between the acquisition of different linguistic skills.

Meta-analytic methods can be applied to any literature, but we believe that developmental research provides a particularly important case where they can contribute to theory development. One reason is that developmental studies may be uniquely vulnerable to false findings because collecting data from children is expensive, and thus sample sizes are often small and studies are underpowered. In addition, the high cost and practical difficulties associated with collecting large developmental datasets means that replications are relatively rare in the field. Meta-analysis provides a method for addressing these issues by harnessing existing data to estimate effect sizes and developmental trends.

We take as our ultimate goal a broad theory of language acquisition that can explain and predict the range of linguistic skills a child acquires. As a first step toward this end, we

collected a dataset of effect sizes in the language acquisition literature across 12 phenomena (Metalab; http://metalab.stanford.edu). We use this dataset to 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 acquisition research. We find broadly that this literature has strong evidential value, and thus that the effects reported in the literature should constrain our theorizing of language acquisition. We then turn toward the task of synthesizing these findings across phenomena and offer a preliminary, quantitative synthesis.

We take as our ultimate goal a broad theory of language acquisition that can explain and predict the range of linguistic skills a child acquires. As a first step toward this end, we collected a dataset of effect sizes in the language acquisition literature across 12 phenomena at many different levels of linguistic representation and processing (Metalab; http://metalab.stanford.edu; see Table 1 for description of phenomena). We use this dataset to demonstrate how meta-analysis supports building theory building in two ways. We first use meta-analytic techniques to evaluate the evidential value of the empirical landscape in language acquisition research. We find broadly that this literature has strong evidential value, and thus that the effects reported in the literature should constrain our theorizing of language acquisition. We then turn toward the task of synthesizing these findings across phenomena and offer a preliminary, quantitative synthesis.

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 (Simonsohn, Nelson, & Simmons, 2014b, 2014a; Simonsohn, Simmons, & Nelson, 2015). These analytical approaches each have limitations, but taken together, they provide converging evidence about whether an effect is likely to exist, and the extent to which publication bias and other questionable research practices are present in the literature.

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 (49)
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.	29 (114)
	Vowel discrimination (non-native) (Tsuji & Cristia, 2014)	Discrimination of non-native vowels, including results from a variety of methods.	15 (48)
	Statistical sound learning (Cristia, in prep.)	Infants' ability to learn sound categories from their acoustic distribution.	9 (17)
	Word segmentation (Bergmann & Cristia, 2015)	Recognition of familiarized words from running, natural speech using behavioral methods.	68 (285)
Words	Mutual exclusivity (Lewis & Frank, in prep.)	Bias to assume that a novel word refers to a novel object in forced-choice paradigms.	20 (60)
	Sound Symbolism (Lammertink et al., 2016)	Bias to assume a non-arbitrary relationship between form and meaning ("bouba-kiki effect") in forced-choice paradigms.	11 (44)
	Concept-label advantage (Lewis & Long, unpublished)	Infants' categorization judgments in the presence and absence of labels.	14 (49)
	Online word recognition (Frank, Lewis, & MacDonald, 2016)	Online word recognition of familiar words using two-alternative forced choice preferential looking.	6 (14)
Communication	Gaze following (Frank, Lewis, & MacDonald, 2016)	Gaze following using standard multi-alternative forced-choice paradigms.	12 (33)
	Pointing and vocabulary (Colonnesi et al., 2010)	Concurrent correlations between pointing and vocabulary.	12 (12)

Table 1
Overview of meta-analyses in dataset.

Overall, we find most phenomena in the language acquisition literature have evidential value, and can therefore provide the basis for theoretical development. We also find evidence for some bias, as well as evidence that two phenomena—phonotactic learning and statistical

sound learning—likely describe null or near-null effects.

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 phenomena. We find evidence for a non-zero effect size in 10 out of 12 of the phenomena in our dataset, suggesting these literatures describe non-zero effects. In the case of phonotactic learning and sound category learning, however, we find that the meta-analytic effect size estimate does not differ from zero, indicating that these literatures do not describe robust effects (as first reported in Cristia, in prep.).

We next turn to methods of assessing evidential value that describe 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 attempt to 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). This number provides an estimate of the size and variance of an effect using the intuitive unit of number of studies. To calculate this effect, we estimated the overall effect size for each phenomenon (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 meta-analyses we assessed, this analysis suggests a very large number of studies would have to be "missing" in each literature (M = 3,470) in order for the overall effect sizes to be 0. Thus, while it is

Phenomenon	***d***	**fail-safe-N**	**funnel skew**	**p-curve skew**
IDS preference	0.7 [0.52, 0.88]	3507	1.5	-10.4*
Phonotactic learning	0.04 [-0.09, 0.16]	45	-1.43	-1.52
Vowel discrim. (native)	0.68 [0.56, 0.81]	8724	8.55*	-9.76*
Vowel discrim. (non-native)	0.66 [0.42, 0.9]	3391	3.86*	-8.89*
Statistical sound learning	-0.19 [-0.42, 0.03]	†	-2.99*	-1.03
Word segmentation	0.19 [0.14, 0.23]	5374	2.59*	-9.4*
Mutual exclusivity	1.01 [0.68, 1.33]	6443	8.26*	-12.87*
Sound symbolism	0.12 [-0.02, 0.25]	526	1.42	-5.56*
Concept-label advantage	0.47 [0.33, 0.61]	2337	1.37	-4.79*
Online word recognition	1.36 [0.84, 1.88]	1934	2.61*	-14.51*
Gaze following	1.27 [0.93, 1.61]	4277	3.3*	-18.66*
Pointing and vocabulary	0.98 [0.62, 1.34]	1617	1.25	-6.33*

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 (Sterne & Egger, 2005); p-curve skew = test of the right skew of the p-curve using the Stouffer method (Simonsohn, Simmons, & Nelson, 2015); Brackets give 95% confidence intervals, and parentheses show p-values. †Fail-safe-N is not available here because the meta-analytic effect size estimate is less than 0.

possible that some reporting bias is present in the literature, the overall large fail-safe-N suggests that the literature nonetheless likely describes robust effects.

This analysis provides a quantitative estimate of the size of an effect in an intuitive unit, but it does not assess analytical or publication bias (Scargle, 2000). Importantly, if experimenters are exercising analytical flexibility through practices like selective reporting of analyses or p-hacking, then the number and magnitude of observed true effects in the literature may be greatly inflated. In the next analysis, we assess the presence of bias through funnel plots.

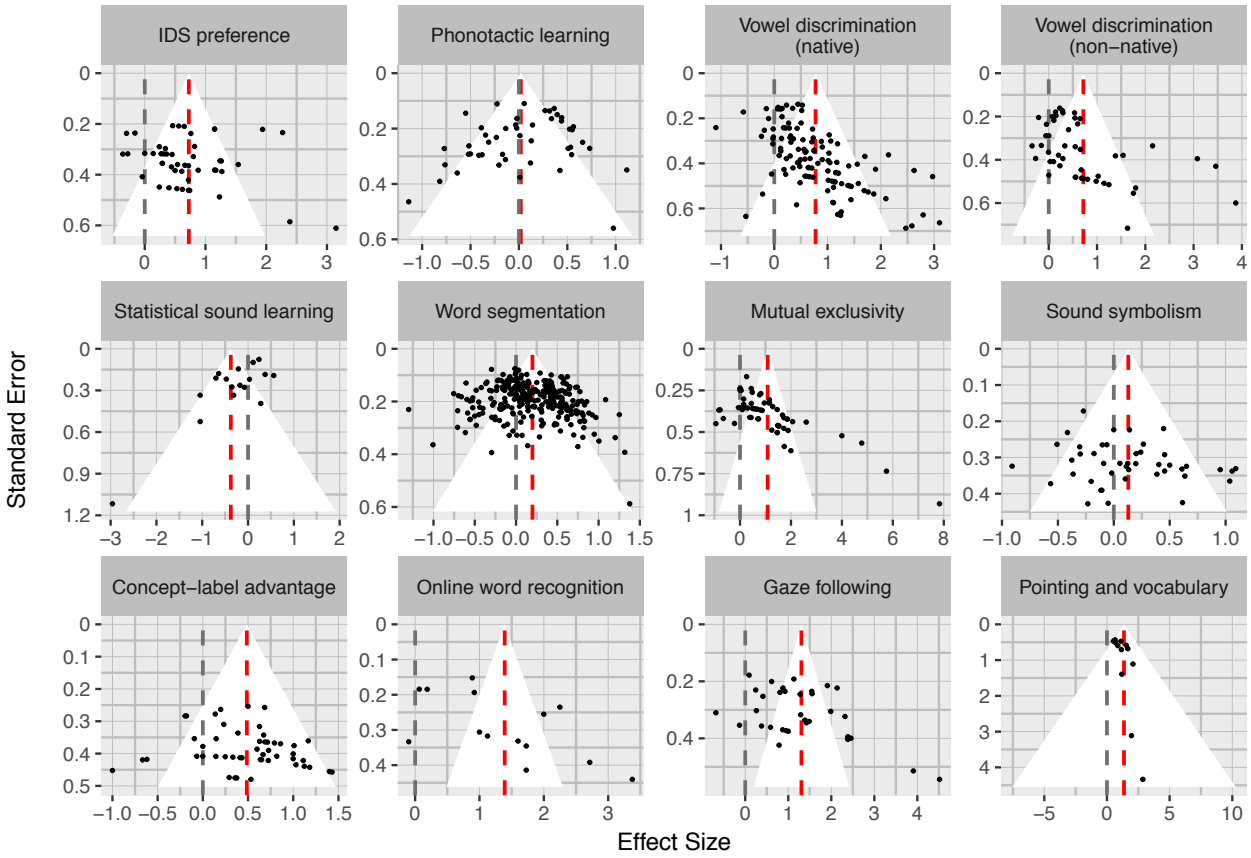


Figure 1. Funnel plots for each meta-analysis. Each effect size estimate is represented by a point, and the mean effect size is shown as a red dashed line. The grey dashed line shows an effect size of zero. The funnel corresponds to a 95% CI around this mean. In the absence of true heterogeneity in effect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

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 phenomena (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 heterogeneity in phenomena (e.g., different ages).[1] P-curve analyses provide one method for addressing this issue, which we turn to next.

P-curves

A p-curve is the distribution of p-values for the statistical test of the main hypothesis across a literature (Simonsohn et al., 2014b, 2014a, 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 heterogeneity 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.

P-values for each condition were calculated based on the reported test statistic. However, test statistics were not available for many conditions, either because they were not reported or because they were not coded. To remedy this, we also calculated p-values indirectly based on descriptive statistics (means and standard deviations; see SI for details).

Figure 2 shows p-curves for each of our 12 meta-analyses. All p-curves show evidence of right skew, with the exception of phonotactic learning and statistical sound learning (Table 2, column 5). This pattern did not differ when only reported test-statistics were used to calculate p-curves (see SI).

In sum, then, meta-analytic methods, along with our dataset of effect sizes, provide an opportunity to assess the replicability of the field of language acquisition. Across a range of analyses, we find that this literature shows some evidence for bias, but overall, it is quite robust.

[1] The role of moderators such as age can be interactively explored on the Metalab website (http://metalab.stanford.edu).

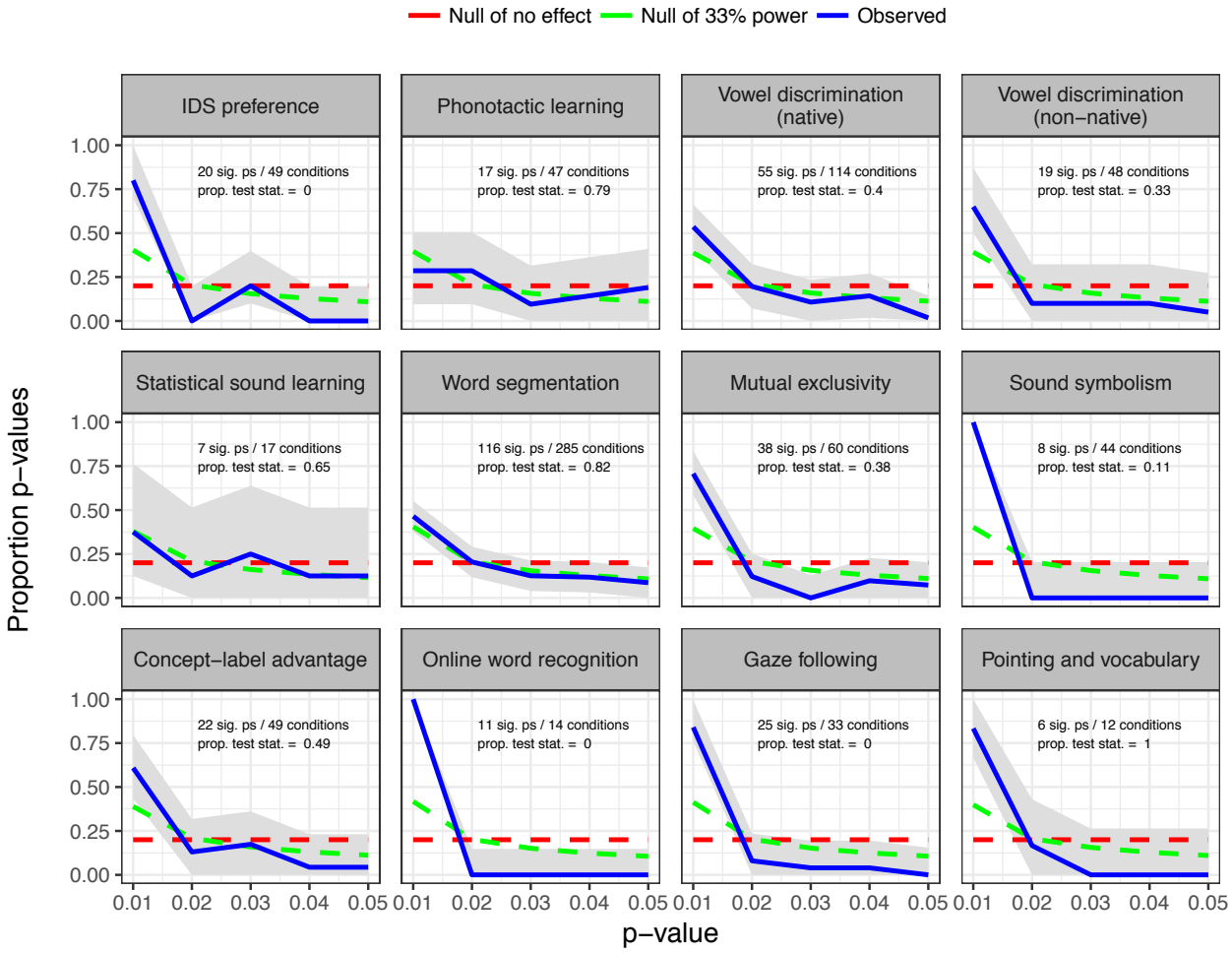


Figure 2. P-curve for each meta-analysis (Simonsohn, Nelson, & Simmons, 2014). In the absence of p-hacking, we should expect the observed p-curve (blue) to be right-skewed (more small values). The red dashed line shows the expected distribution of p-values when the effect is non-existent (the null is true). The green dashed line shows the expected distribution if the effect is real, but studies only have 33% power. Grey ribbons show 95% confidence intervals estimated from a multinomial distribution. Text on each plot shows the number of p-values for each dataset that are less than .05 and thus are represented in each p-curve ("sig. ps"), relative to the total number of conditions for that phenomenon. Each plot also shows the proportion of p-values that were derived from test statistics reported in the paper ("prop. test stat."); all others were derived by conducting analyses on the descriptive statistics or transforming reported effect sizes.

Quantitative Evaluation of Theories

Next, we turn to how these data can be used to constrain and develop theories of language acquisition.

Meta-analytic methods provide a precise, quantitative description of the developmental trajectory of individual phenomena. Figure 3 presents the developmental trajectories of the phenomena in our dataset at each level in the linguistic hierarchy. By describing how effect sizes change as a function of age, we can begin to understand what factors might moderate that trajectory, such as aspects of a child's experience or maturation. For example, the meta-analysis on mutual exclusivity (the bias for children to select a novel object, given a novel word; Markman & Wachtel, 1988) suggests a steep developmental trajectory of this skill. We then can use these data to build quantitative models to understand how aspects of experience (e.g., vocabulary development) or maturational constraints may be related to this trajectory (e.g., Frank, Goodman, & Tenenbaum, 2009; McMurray, Horst, & Samuelson, 2012).

In addition, meta-analytic methods provide an approach for synthesizing across different linguistic skills via the language of effect sizes. The ultimate goal is to use meta-analytic data to build a single, quantitative model of the language acquisition system, much like those developed for individual language acquisition phenomena, like word learning. Developing a single quantitative model is a lofty goal, however, and will likely require much more precise description of the phenomena than is available in our dataset. Nevertheless, we can use our data to distinguish between broad meta-theories about the interdependency of skills.

We first consider two intuitive theories of task-to-task dependencies that have been articulated in a number of forms. The stage-like theory proposes that linguistic skills are acquired sequentially beginning with skills at the lowest level of the linguistic hierarchy. Under this theory, once a skill is mastered, it can be used to support the acquisition of skills higher in the linguistic hierarchy. In this way, a child sequentially acquires the skills of

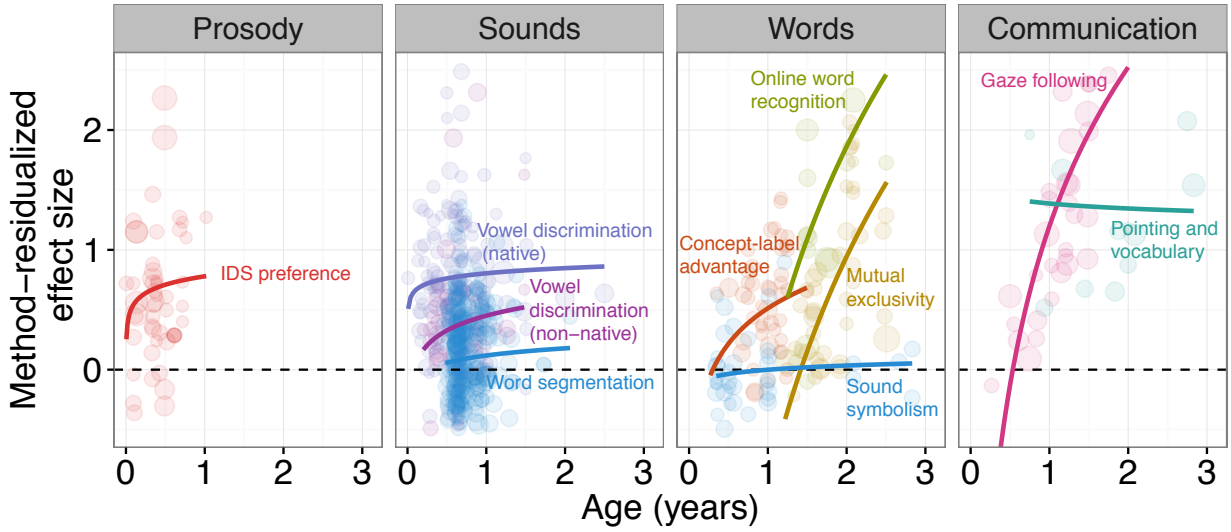


Figure 3. Method-residualized effect size plotted as a function of age across the 10 meta-analyses in our dataset shown to have evidential value (excluding phonotactic learning and sound category learning). Lines show logarithmic model fits. Each point corresponds to a condition, with the size of the point indicating the number of participants.

language, "bootstrapping" from existing knowledge at lower levels to new knowledge at higher levels. There is a wide range of evidence consistent with this view. For example, there is evidence that prosody supports the acquisition of sound categories (e.g., Werker et al., 2007), word boundaries (e.g., Jusczyk, Houston, & Newsome, 1999), grammatical categories (e.g., Shi, Werker, & Morgan, 1999), and even word learning (e.g., Shukla et al., 2011).

A second possibility is that there is interactivity in the language system such that multiple skills are learned simultaneously across the system. For example, under this proposal, a child does not wait to begin learning the meanings of words until the sounds of a language are mastered; rather, the child is jointly solving the problem of word learning in concert with other language skills. This possibility is consistent with predictions of a class of hierarchical Bayesian models that suggest that more abstract knowledge may be acquired quickly, before lower-level information, and may in turn support the acquisition of lower information ("blessing of abstraction," Goodman, Ullman, & Tenenbaum, 2011). There is evidence for this proposal from work that suggests word learning supports the acquisition of lower-level information like phonemes (Feldman et al., 2013). More broadly, there is evidence

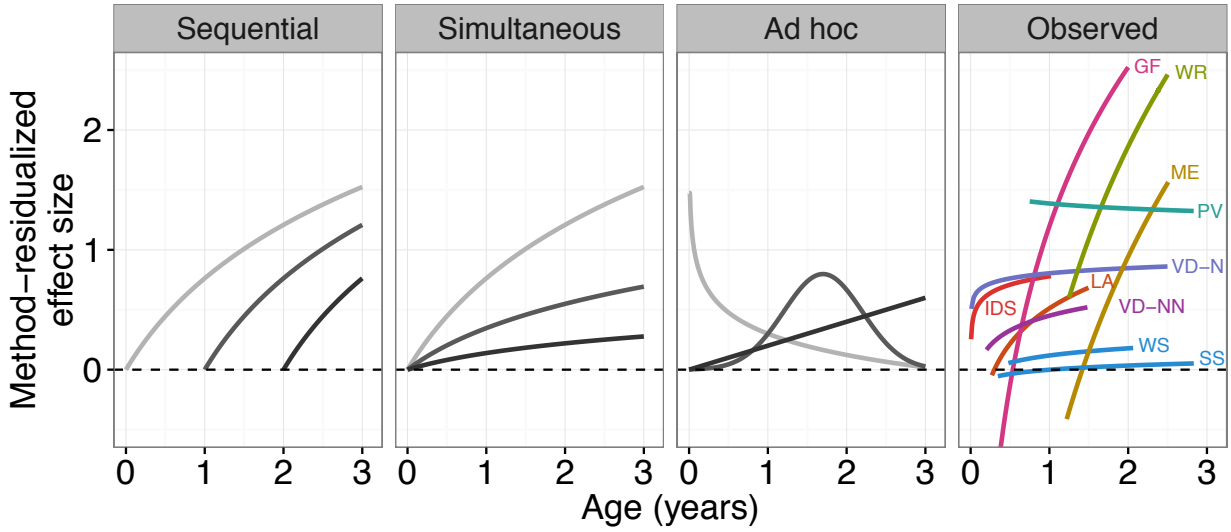


Figure 4. The left two panels show the developmental trajectories predicted under different meta-theories of language acquisition. The stage-like theory predicts that a child will not begin learning the next skill in the linguistic hierarchy until the previous skill has been mastered. The interactive theory predicts that multiple skills may be simultaneously acquired. The third panel shows other possible developmental trajectories (decreasing, linear, and non-monotonic). The fourth panel shows the observed meta-analytic data. Effect size is plotted as a function of age from 0-3 years, across 10 different phenomena (excluding phonotactic learning and sound category learning). Model fits are the same as in Figure 3. These developmental curves suggest there is interactivity across language skills, rather than stage-like learning of the linguistic hierarchy. GF: Gaze following; IDS: IDS preference; LA: Concept -label advantage; ME: Mutual exclusivity; VD-(N)N: Vowel discrimination (non-)native; PV: Pointing-vocabulary correlations; SS: Sound symbolism; WR: Word recognition.

that higher-level skills like word learning may be acquired relatively early in development, likely before lower level skills have been mastered (e.g., Bergelson & Swingley, 2012; Tincoff & Jusczyk, 1999).

These two theories make different predictions about relative trajectories of skills across development. Within the meta-analytic framework, we can represent these different trajectories schematically by plotting the effect sizes for different skills across development. In particular, the bottom-up theory predicts serial acquisition of skills (Figure 4; left) while the interactive theory predicts simultaneous acquisition (left center). We can also specify many other possible trajectories by varying the functional form and parameters of the model. Figure 4 ("Ad hoc"; right center) shows several other possible trajectories. For example, a

skill might have a non-monotonic trajectory, increasing with age, and then decreasing. By specifying the shape of these developmental trajectories and the age at which acquisition begins, we can consider many patterns of developmental trajectories, and how these different patterns, in turn, constrain our meta-theories of development.

Our data allow us to begin to differentiate between this space of theories. Figure 4 (right) presents a synthetic representation of the developmental trajectories of the skills in our dataset with literatures shown to have evidential value (all but phonotactic learning and sound category learning). We find strong evidence for the simultaneous acquisition of skills—children begin learning even high-level skills, like the meanings of words, early in development, and even low-level skills like sound categories show a protracted period of development. This pattern is consistent with an interactive theory of language acquisition, and at least prima facie inconsistent with stage-like theories. In future research, we can use this approach to distinguish between a larger space of meta-theories and, ultimately, refine our way towards a single quantitative theory of language acquisition.

Discussion

Building a theory of a complex psychological phenomenon requires making good inductive inferences from the available data. Meta-analysis can support this process by providing a toolkit for quantitative description of individual behaviors and their relationship to important moderators (e.g., age, in our case). Here, we apply the meta-analytic toolkit to the domain of language acquisition—a domain where there are concerns of replicability, and where high-fidelity data are needed for theory building. We find that the existing literature in this domain describes mostly robust phenomena and thus should form the basis of theory development. We then aggregate across phenomena to offer the first quantitative synthesis of the field. We find evidence that linguistic skills are acquired interactively rather than in a stage-like fashion.

In this paper, we focused on theoretical motivations for building meta-analysis, but

naturally, there are many other practical reasons for conducting a quantitative synthesis. For example, 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 or measure tends to yield overall larger effect sizes than another, the strategic researcher might select this paradigm in order to maximize the power achieved with a given sample size. These and other advantages, illustrated with the same database used here, are explained in Bergmann et al. (in prep.).

Despite its potential, there are a number of important limitations to the meta-analytic method as tool for theory building in psychological research. One challenging issue is that in many cases method and phenomenon are confounded. This is problematic because a method with less noise than another will produce a bigger effect size for the same phenomenon. As a result, it is difficult to determine the extent to which a difference in effect size between two phenomena is due to an underlying difference in the phenomena, or merely to a difference in they way it was tested. While method may account for some variability in our dataset, we find that method does not have a large impact on effect size for phenomena, relative to other moderators like age (see SI). Nevertheless, the covariance between method and phenomenon in our dataset limits our ability to directly compare effect sizes across phenomena.

Second, meta-analysis, like all analysis methods, requires the researcher to make analytical decisions, and these decisions may be subject to the biases of the researcher. We believe that a virtue of the current approach is that we have applied the same analytical method across all phenomena we examined, thus limiting our "degrees of freedom" in the analysis. However, in some cases this uniform approach to data analysis means that we are unable to take into consideration aspects of a particular phenomenon that might be relevant. For example, in a stand-alone meta-analysis on vowel discrimination, Tsuji and Cristia (2014) elected only to include papers that tested at least two different age groups as a way of focusing on age differences while controlling for other possible differences between

experiments. Others however might have reasonably dealt with this issue in another way, by normalizing effect sizes across methods, for example. Notably, this analytical decision has consequences for interpretation: Tsuji and Cristia (2014) found a moderate decrease in effect size with age for non-native vowel discrimination, while the current analysis suggests a moderate increase. We believe that the systematic, uniform analytical approach used here is the most likely to minimize bias by the researcher and reveal robust psychological phenomena. There may be cases however where this one-size-fits-all approach is inappropriate, particularly in meta-analyses with high heterogeneity.

There are also limits to this method for inferring a meta-theory of language acquisition. Meta-theories of language acquisition suggest a particular causal relationship between different skills and how they change over development. For example, the interactive theory suggests that skills at higher levels *support* the acquisition at lower levels, even before skills at lower levels are mastered. In the meta-analytic framework, this predicts that there should be simultaneous development of skills across the language hierarchy—as we observe in the current work. Importantly, however, this analysis is inherently correlational, and therefore we cannot directly infer a causal relationship between acquisition at lower levels and acquisition at higher levels. That is, while the observed pattern is consistent with the interactive theory, it is also possible that there is no causal relationship between skills across the language hierarchy, merely parallel trajectories of acquisition. For this reason, experimental work must go hand-in-hand with meta-analysis to address causal questions.

Finally, there are a number of important limitations to the meta-analytic method more broadly. One issue is that the method relies on researchers conducting replications of the same study across a range of ages and, critically, reporting these data so that they can be used in meta-analyses. To the extent that researchers do not conduct these studies, or report the necessary statistics in their write-ups (e.g., means and standard deviations), the meta-analytic method cannot be applied. In addition, the meta-analytic method, as in the case of qualitative forms of synthesis (e.g., literature review), is limited by the potential

presence of bias, which can come from a range of sources including non-representative participant populations, failure to publish null findings, and analytical degrees-of-freedom. To the extent these biases are present in the literature, methods of synthesizing these findings will also be biased.

In sum, understanding the psychological mechanisms underlying complex phenomena is a difficult inferential task: The researcher must develop a predictive and explanatory theory on the basis of limited and noisy experimental data. Here we have focused on language acquisition as a case study of how meta-analytic methods can be productively leveraged as a tool for theory building. Meta-analytic methods allow the researcher to determine whether phenomena are robust, synthesize across contradictory findings, and ultimately, build an integrative theory across phenomena. Moving forward, we see meta-analysis as a powerful tool in the researcher's toolkit for developing quantitative theories to account for complex psychological phenomena.

Methods

We analyzed 12 different phenomena in language acquisition. We selected these particular phenomena because of their theoretical importance or because a previously-published meta-analysis already existed.

To obtain estimates of effect size, we either coded or adapted others' coding of papers reporting experimental data (see SI for details). Within each paper, we calculated a separate effect size estimate for each experiment and age group (we refer to each measurement separated by age as a "condition"). In total, our sample includes estimates from 227 papers, 772 different conditions and 9,329 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 for specific search strategies). Nevertheless, meta-analytic methods for aggregating even the smallest sample of studies are likely to be less biased than qualitative methods (Valentine, Pigott, & Rothstein, 2010).

Data and Code Availability. The data and code reported in this paper have been deposited in GitHub, a web-based repository hosting service, https://github.com/langcog/metalab/.

Supplementary Information. This article contains supporting information online at http://rpubs.com/mll/synthesisSI

Author Contributions. ML, ST, CB, PP, AC, and MF wrote the paper. ML, ST, CB, AC, and MF coded papers for the meta-analytic dataset. All authors contributed to data analysis. MB, MF, and ML developed the Metalab website infrastructure.

References. Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... others. (2016). Response to comment on "estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bergmann, C., & Cristia, A. (2015). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *19*(6), 901–917.

Bergmann, C., Tsuji, S., Piccinini, P., Lewis, M., Braginsky, M., Frank, M., & Cristia, A. (in prep). Building broad-shouldered giants: Synthesizing studies to plan for reproducible research.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Colonnesi, C., Stams, G. J. J., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, *30*(4), 352–366.

Cristia, A. (in prep). Infants' phonology learning in the lab.

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

Ebersole, C., Atherton, O., Belanger, A., Skulborstad, H., Adams, R., Allen, J., & Nosek, B. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013).

Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438.

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

Frank, M. C., Lewis, M., & MacDonald, K. (2016). A performance model for early word learning. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on Estimating the reproducibility of psychological science. *Science*, *351*(6277), 1037–1037. doi:10.1126/science.aad7243

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Johnson, M., Demuth, K., Jones, B., & Black, M. J. (2010). Synergies in learning words and their referents. In *Advances in neural information processing systems* (pp. 1018–1026).

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive Psychology*, *39*(3), 159–207.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843.

Lammertink, I., Fort, M., Peperkamp, S., Fikkert, P., Guevara-Rukoz, A., & Tsuji, S. (2016).

SymBuki: A meta-analysis on the sound-symbolic bouba-kiki effect in infants and toddlers. Poster presented at the XXI Biennial International Congress of Infant Studies, New Orleans, USA.

Lewis, M. L., & Frank, M. C. (in prep). Mutual exclusivity: A meta-analysis.

Lewis, M., & Long, B. (unpublished). Meta-analysis of the concept-label advantage.

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

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.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

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

Scargle, J. D. (2000). Publication bias: The "file-drawer problem" in scientific inference. *Journal of Scientific Exploration*, *14*(1), 91–106.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*(2), B11–B21.

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the*

National Academy of Sciences, *108*(15), 6038–6043.

Simonsohn, Nelson, L. D., & Simmons, J. P. (2014a). 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. (2014b). 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. *Journal of Experimental Psychology: General*, *144*(6), 1146–52.

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.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175.

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

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147–162.

A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Supplementary Information

Molly Lewis, Mika Braginsky, Sho Tsuji, Christina Bergmann, Page Piccinini, Alejandrina Cristia, and Michael C. Frank

2017-01-24

Contents

This document was created from an R markdown file. Data from the paper can be interactively explored on the Metalab website, http://metalab.stanford.edu/. The manuscript itself was also produced from an R markdown file, and thus all analyses presented in the paper can be reproduced from that document. Supplementary materials can also be viewed online: http://rpubs.com/mll/synthesisSI.

Search strategies

Meta-analyses were conducted by the authors for all but two phenomena (IDS preference and Pointing and Vocabulary). Data for these two phenomena were obtained by adapting effect size estimates from existing, published meta-analyses (Dunst, Gorman, & Hamby, 2012; Colonnesi et al., 2010). Across phenomena, meta-analyses varied in their degree of systematicty in selecting papers. In the table below, we describe the search strategy for each phenomenon. Quoted descriptions are taken directly from the published source.

```
methods.table = datasets %>% select(name, search_strategy,
    internal_citation) %>% mutate(name = as.factor(name),
    name = plyr::revalue(name, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Native vowel discrimination",
        `Vowel discrimination (non-native)` = "Non-native vowel discrimination"))) %>%
    mutate(name = paste(name, internal_citation)) %>%
    select(-internal_citation) %>% .[c(1, 6, 4, 5,
    7, 8, 10, 12, 2, 9, 3, 11), ] %>% rename(Phenomenon = name,
    `Search Strategy` = search_strategy)
```

Phenomenon	Search Strategy
IDS preference (Dunst, Gorman, & Hamby, 2012)	"Studies were located using motherese or parentese or fatherese or infant directed speech or infant-directed speech or infant directed talk or child directed speech or child-directed speech or child directed talk or child-directed talk or baby talk AND infant* or neonate* or toddler* as search terms. Both controlled-vocabulary and natural-language searches were conducted (Lucas & Cutspec, 2007). Psychological Abstracts (PsychInfo), Educational Resource Information Center (ERIC), MEDLINE, Academic Search Premier, CINAHL, Education Resource Complete, and Dissertation Abstracts International were searched. These were supplemented by Google Scholar, Scirus, and Ingenta searches as well as a search of an extensive EndNote Library maintained by our Institute. Hand searches of the reference sections of all retrieved journal articles, book chapters, books, dissertations, and unpublished papers were also examined to locate additional studies. Studies were included if the effects of infant-directed speech on child behavior were compared to the effects of adult-directed speech on child behavior. Studies that intentionally manipulated word boundaries (e.g., Hirsh-Pasek et al., 1987; Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989) or used nonsense words or phrases (e.g., Mattys, Jusczyk, Luce, & Morgan, 1999; Thiessen, Hill, & Saffran, 2005) were excluded."
Phonotactic learning (Cristia, in prep.)	"Studies were considered based on a forward search from the seminal paper (MWG02) on both pubmed and google search, a list of references produced by the author, direct contact of labs having published on the topic, and announcements to several mailing lists."
Native vowel discrimination (Tsuji & Cristia, 2014)	"A full search on scholar.google.com was conducted in September 2012 with the keyword combination "{infant|infancy} & {vowel|speech sound|syllable} & discrimination." Additionally, the search terms were translated into French, German, Japanese, and Spanish for additional searches. We also asked experts in the field to inform us of any published or unpublished studies we had missed. Experts were defined as scientists having participated in at least two studies identified in our intermediate search sample or who were part of a lab where such research had taken place, and who were still active in the field or could be otherwise contacted. Further, articles were added based on a screening of articles cited and articles citing the articles in the remaining search sample. The complete sample is available as a public resource (Tsuji & Cristia, in preparation, https://sites.google.com/site/inphondb/). The search sample was then narrowed down to the final search sample of 19 articles." (See paper for additional details)
Non-native vowel discrimination (Tsuji & Cristia, 2014)	See Native Vowel Discrimination, above.
Statistical sound learning (Cristia, in prep.)	"Studies were considered based on a forward search from the seminal paper (MWG02) on both pubmed and google search, a list of references produced by the author, manual inspection of several editions of two leading conferences (ISIS and IASCL 2004-2012), direct contact of labs having published on the topic, and announcements to several mailing lists."
Word segmentation (Bergmann & Cristia, 2015)	"We first generated a list of potentially relevant items to be included in our meta-analysis using the Google Scholar search engine, with the broad search term 'infant word segmentation' (following Gehanno, Rollin & Darmoni, 2013). This search was carried out on 27 November 2012 and we inspected the first 1000 results. Fifteen additional items were included based on recommendations and by scanning references of included papers. After removing duplicates, we screened the title and abstract of each remaining item and identified 231 items for full-text inspection using the following inclusion criteria: (1) original data were reported; (2) the stimulus material was continuous natural speech spoken in the participants' native language; (3) the dependent measure was looking time (LT) at a neutral visual target (i.e. not a possible referent of one set of stimuli); (4) infants were normally developing."

Mutual exclusivity (Lewis & Frank, in prep.)	"We conducted a forward search based on citations of Markman and Wachtel (1988) in Google Scholar (September 2013). We also searched from papers using the keyword combination "mutual exclusivity" in both PsychInfo and Google Scholar. We identified additional papers that were cited from this initial list. From these, we identified a relevant subset using the following criteria: (a) monolingual child participants, (b) one familiar object present at test, (c) referents were objects (not facts or object parts), (d) no incongruent cues (e.g. eye gaze at familiar object), and (e) peer-reviewed. We also included a series of studies reported in Frank et al. (2016)."
Sound symbolism (Lammertink et al., 2016)	"We followed the PRISMA statement (Moher, Liberati, Tetzlaff, Altmann, & The PRISMA Group, 2009) for selecting and reporting on the studies to be included in our meta-analysis. We decided to include articles if they were assessing the on-line processing of sound-symbolically matching or mismatching sound-shape correspondences related to the bouba-kiki effect (thus, testing both 'round' and 'spiky' correspondences) in children up to and inclusive of the age of 3 years. 'Matching' responses refer to children's responses to congruent sound-shape associations (round word+round object; spiky word+spiky object) and 'mismatching' responses refer to children's responses to incongruent sound-shape associations (round word+spiky shape; spiky word+round shape), respectively. Since we were already aware of 10 published articles, conference presentations or conference proceedings papers that fit our inclusion criteria,and since we considered our strict inclusion this criteria to lead to a rather small selection of articles, we chose a seed strategy rather than a broad literature search. We began by assembling 4 key articles that fit the inclusion criteria (Asano et al., 2015; Maurer, Pathman & Mondloch, 2006; Ozturk, Krehm & Vouloumanos, 2012; Spector & Maurer, 2013) as well as two recent review papers on sound symbolism including infancy (Imai & Kita, 2014; Lockwood & Dingemanse, 2015). For all of these articles, we screened all potentially relevant references cited in these articles as well as references citing these articles and 'related articles' on their title and abstracts on scholar.google.com. Additionally, we screened titles and abstracts of all articles that cited one of the 4 key articles mentioned above (Asano et al. 2015: 16 citations; Maurer et al. 2006: 241 citations; Ozturk et al. 2012: 54 citations and Spector and Maurer, 2013: 9 citations). This search did not lead to additional eligible articles. In addition, we were aware of 9 conference presentations or conference proceedings papers that fit our inclusion criteria and were not redundant with one of our seed articles, including three by the two first authors of the present article."
Concept-label advantage (Lewis & Long, unpublished)	"We conducted a forward search based on Balaban and Waxman (1997) in Google Scholar and Web of Science (October 2015). This was supplemented with papers identified through citations and publication lists on lab websites. The final sample included only peer-reviewed publications."
Online word recognition (Frank, Lewis, & MacDonald, 2016)	"We conducted a systematic literature review by using Google Scholar to identify peer-reviewed papers citing Fernald et al. (1998). We screened this sample manually to find the subsample of 12 papers that reported both accuracy and reaction time with sufficient detail to permit coding."
Gaze following (Frank, Lewis, & MacDonald, 2016)	"We identified papers using a Google Scholar search for "gaze following" and included those studies that (a) included data from typically-developing children, (b) used a standard face-to-face gaze-following task, and (c) reported percentage accuracy (rather than a score or other composite measure). Although we coded all papers that fit these criteria, we focused on papers with a simple two-alternative forced choice (9 papers); integrating across different numbers of alternatives added additional complexity to our model. In our first iteration of this analysis, we found that very few studies reported reaction times for gaze following, and those that did had no data from children older than 15 months and no data from gaze plus pointing. To remedy this issue we include new analyses of data from Yurovsky, Wade, & Frank (2013) and Yurovsky & Frank (2015)."

Pointing and vocabulary (Colonnesi et al., 2010)	"The search method involved inspection of digital databases (Web of Knowledge, Picarta, PsychInfo) using the following keywords: pointing, gesture, declarative, imperative, precursors, language, words, vocabulary, infancy, intentional communication, and joint attention. Inspection of the reference section of relevant literature was an additional search method (ancestry method). Additionally, also unpublished sources were consulted, such as dissertations and presentations and studies under revision, by using Google Scholar, contacting researchers in the field and consulting digital databases of dissertations (e.g., PROQUEST). Three selection criteria were used to select studies: (a) measurement of infant production and/or comprehension of the pointing gesture; (b) measurement of language by assessing either receptive or expressive language; (c) report of a relation between pointing and language or the presence of data in the article allowing the calculation of a relation between pointing and language development. Exclusion criteria were: (a) subjects with mental or developmental disorders; (b) children older than 60 months; (c) studies in which the pointing gesture was not coded separately from other gestures."

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 and F). Several pre-existing MAs deal with special cases, and these are listed in the script. Except where noted, formulas are from Hedges & Olkin's textbook (2014). All analyses were conducted with the metafor package (Viechtbauer, 2010), using random-effects models. Note that a subset of individual MAs (such as Sound Symbolism) contain effect sizes that are not statistically independent, while the current implementation of random-effect models assumes independence of all individual effect sizes.

For many analyses, the use of a multi-level approach (with grouping by paper) is useful. We do not implement these models in our main analyses because many common statistics are not implemented for these models, e.g. the test for funnel-plot asymmetry. The table below compares the overall ES estimate for the multi-level model to the random effect model (presented in the main text). 95% confidence intervals are given in brackets. These models differ only slightly in their estimates of overall effect size, and in no case do they affect whether the ES differs from zero.

```
overall_es <- function(ma_data, multilevel) {
    if (multilevel) {
        model = metafor::rma.mv(d_calc, V = d_var_calc,
            random = ~1 | short_cite, data = ma_data)
    } else {
        model = metafor::rma(d_calc, d_var_calc, data = ma_data)
    }
    data.frame(dataset = ma_data$short_name[1], overall.d = model$b,
        ci_lower = model$ci.lb, ci_upper = model$ci.ub)
}

all_ds_random = all_data %>% split(.$short_name) %>%
    map(function(ma_data) overall_es(ma_data, 0)) %>%
    bind_rows() %>% mutate_each_(funs(round(., digits = 2)),
    vars = c("overall.d", "ci_lower", "ci_upper")) %>%
    mutate(d_string_random = paste0(overall.d, " [",
        ci_lower, ", ", ci_upper, "]")) %>% mutate(short_name = dataset) %>%
    select(short_name, d_string_random)

all_ds_multi = all_data %>% split(.$short_name) %>%
    map(function(ma_data) overall_es(ma_data, 1)) %>%
```

```
    bind_rows() %>% mutate_each_(funs(round(., digits = 2)),
    vars = c("overall.d", "ci_lower", "ci_upper")) %>%
    mutate(d_string_multi = paste0(overall.d, " [",
        ci_lower, ", ", ci_upper, "]")) %>% mutate(short_name = dataset) %>%
    select(short_name, d_string_multi)

left_join(all_ds_random, all_ds_multi) %>% left_join(select(datasets,
    name, short_name)) %>% select(name, d_string_random,
    d_string_multi) %>% mutate(name = as.factor(name),
    name = plyr::revalue(name, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Native vowel discrimination",
        `Vowel discrimination (non-native)` = "Non-native vowel discrimination"))) %>%
    .[c(2, 8, 3, 4, 10, 5, 7, 11, 6, 12, 1, 9), ] %>%
    kable(col.names = c("Phenomenon", "Random-effect model ES",
        "Mixed-effect model ES"), row.names = F, align = c("l",
        "r", "r"))
```

Phenomenon	Random-effect model ES	Mixed-effect model ES
IDS preference	0.7 [0.52, 0.88]	0.74 [0.47, 1.01]
Phonotactic learning	0.04 [-0.09, 0.16]	0.12 [-0.01, 0.25]
Native vowel discrimination	0.68 [0.56, 0.81]	0.7 [0.51, 0.89]
Non-native vowel discrimination	0.66 [0.42, 0.9]	1 [0.41, 1.59]
Statistical sound learning	-0.19 [-0.42, 0.03]	-0.26 [-0.58, 0.05]
Word segmentation	0.19 [0.14, 0.23]	0.16 [0.11, 0.21]
Mutual exclusivity	1.01 [0.68, 1.33]	0.82 [0.54, 1.1]
Sound symbolism	0.12 [-0.02, 0.25]	0.22 [0, 0.44]
Concept-label advantage	0.47 [0.33, 0.61]	0.41 [0.26, 0.57]
Online word recognition	1.36 [0.84, 1.88]	1.24 [0.74, 1.74]
Gaze following	1.27 [0.93, 1.61]	1.17 [0.8, 1.55]
Pointing and vocabulary	0.98 [0.62, 1.34]	0.98 [0.62, 1.34]

Funnel plots

```
funnel.es.with.outliers = all_data %>% mutate(dataset = as.factor(dataset),
    dataset = gdata::reorder.factor(dataset, new.order = c(2,
        6, 10, 11, 9, 12, 4, 8, 3, 5, 1, 7)), dataset = plyr::revalue(dataset,
        c(`Infant directed speech preference` = "IDS preference",
            `Statistical sound category learning` = "Statistical sound learning",
            `Label advantage in concept learning` = "Concept-label advantage",
            `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
            `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)"))) %>%
    group_by(dataset) %>% mutate(outlier = ifelse(d_calc >
    mean(d_calc) + (3 * sd(d_calc)) | d_calc < mean(d_calc) -
    (3 * sd(d_calc)), 1, 0), outlier = as.factor(outlier))
```

If an effect size is an extreme outlier from the overall mean, this may indicate that the effect size estimates a different psychological phenomenon than the one estimated by others in the sample. One approach for dealing with this type of heterogenity is to exclude outliers from analyses. Here we present funnel plots that

exclude effect sizes that lie 3 standard deviations above or below the mean effect size for each meta-analysis. Of the 772 effect sizes in the dataset, 7 were outliers (0.9%).

```
CRIT_95 = 1.96

funnel.es.data = funnel.es.with.outliers %>% filter(outlier ==
    0) %>% mutate(se = sqrt(d_var_calc), es = d_calc,
    center = mean(d_calc), lower_lim = max(se) + 0.05 *
        max(se))

# separate df for 95 CI funnel shape
funnel95.data.wide <- funnel.es.data %>% select(center,
    lower_lim, dataset) %>% group_by(dataset) %>% summarise(x1 = (center -
    lower_lim * CRIT_95)[1], x2 = center[1], x3 = center[1] +
    lower_lim[1] * CRIT_95, y1 = -lower_lim[1], y2 = 0,
    y3 = -lower_lim[1])

funnel95.data.x = funnel95.data.wide %>% select(dataset,
    dplyr::contains("x")) %>% gather("coordx", "x",
    2:4) %>% arrange(dataset, coordx) %>% select(-coordx)

funnel95.data.y = funnel95.data.wide %>% select(dataset,
    dplyr::contains("y")) %>% gather("coordy", "y",
    2:4) %>% arrange(dataset, coordy) %>% select(-coordy)

funnel95.data = bind_cols(funnel95.data.x, funnel95.data.y)

ggplot(funnel.es.data, aes(x = es, y = -se)) + facet_wrap(~dataset,
    scales = "free") + xlab("Effect Size") + ylab("Standard Error\n") +
    scale_colour_solarized(name = "") + geom_polygon(aes(x = x,
    y = y), data = funnel95.data, fill = "white") +
    geom_vline(aes(xintercept = x2), linetype = "dashed",
        color = "red", size = 0.8, data = funnel95.data.wide) +
    geom_vline(xintercept = 0, linetype = "dashed",
        color = "grey44", size = 0.8) + scale_y_continuous(labels = function(x) {
    abs(x)
}) + geom_point(size = 0.5) + theme(panel.grid.major = element_line(colour = "grey",
    size = 0.2), panel.grid.minor = element_line(colour = "grey",
    size = 0.5), strip.text.x = element_text(size = 9),
    strip.background = element_rect(fill = "grey"))
```

We next compare the results of funnel skew (Egger's test) for the dataset with outliers excluded to the full dataset (which is reported in the main text). There is a difference in significance for only Statistical Sound Learning: With outliers excluded, these meta-analyses no longer show evidence for skew.

```
eggers_tests <- function(ma_data){
    model = metafor::rma(d_calc, d_var_calc, data = ma_data) # model
    egg.random = metafor::regtest(model) # Egger's test
    data.frame(dataset = ma_data$short_name[1],
               egg.random.z = egg.random$zval,
               egg.random.p = egg.random$pval)
}

eggers.data.f = all_data %>%
```

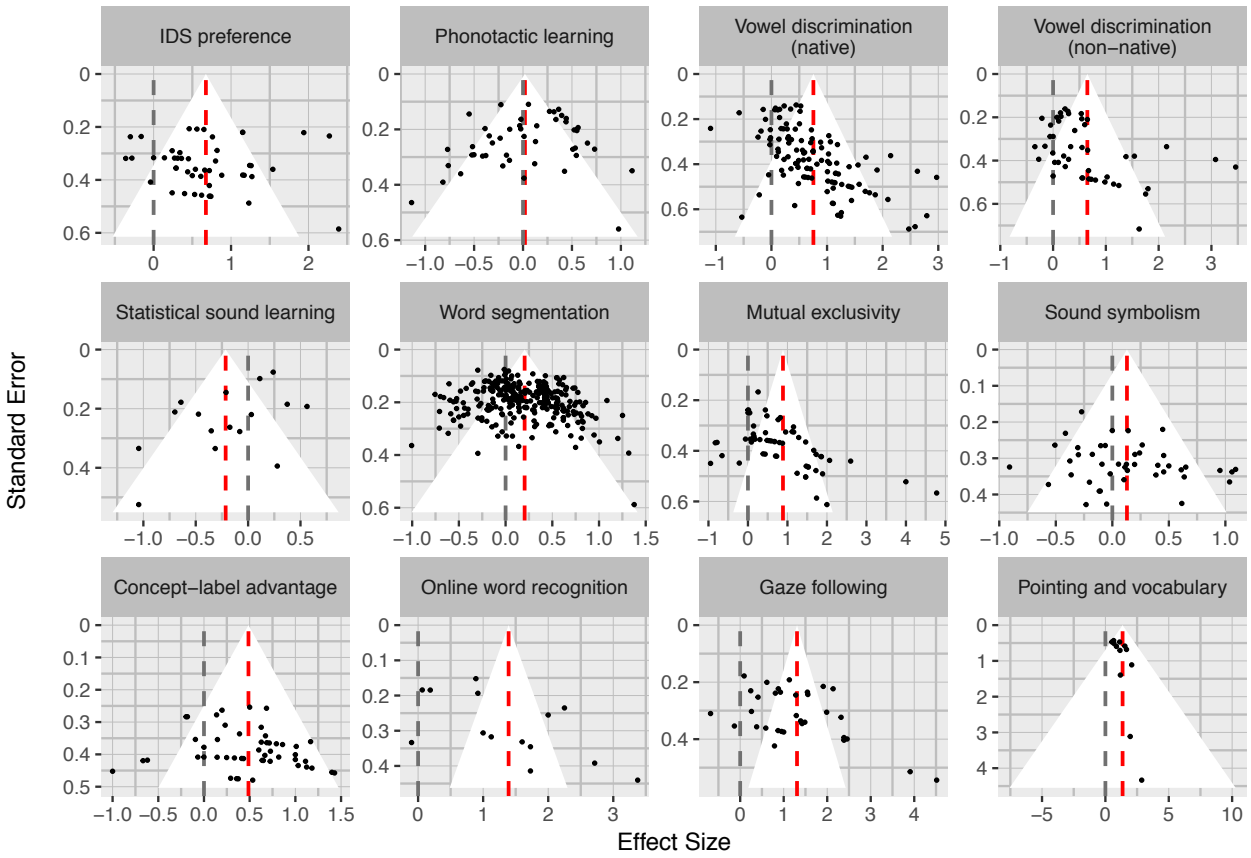


Figure S1: Funnel plots for each meta-analysis with outliers excluded. Each effect size estimate is represented by a point, and the mean effect size is shown as a red dashed line. The grey dashed line shows an effect size of zero. The funnel corresponds to a 95% CI around this mean.

```
  group_by(dataset) %>%
  mutate(outlier = ifelse(d_calc > mean(d_calc)  + (3 * sd(d_calc)) |
                            d_calc < mean(d_calc)  - (3 * sd(d_calc)), 1, 0),
         outlier = as.factor(outlier)) %>%
  filter(outlier == 0) %>%
  ungroup() %>%
  split(.$short_name) %>%
  map(function(ma_data) eggers_tests(ma_data)) %>%
  bind_rows() %>%
  mutate(egg.random.z = round(egg.random.z, digits = 2)) %>%
  mutate(egg.random.p = round(egg.random.p, digits = 2)) %>%
  mutate(egg_string.f = paste0(egg.random.z, " (", egg.random.p, ")")) %>%
  select(dataset, egg_string.f)

eggers.data.all = all_data %>%
  group_by(dataset) %>%
  ungroup() %>%
  split(.$short_name) %>%
  map(function(ma_data) eggers_tests(ma_data)) %>%
  bind_rows() %>%
  mutate(egg.random.z = round(egg.random.z, digits = 2)) %>%
  mutate(egg.random.p = round(egg.random.p, digits = 2)) %>%
  mutate(egg_string.all = paste0(egg.random.z, " (", egg.random.p, ")")) %>%
  select(dataset, egg_string.all)

left_join(eggers.data.all, eggers.data.f) %>%
              ungroup() %>%
             .[c(2,8,3,4,10,5,7,11,6,12,1,9),]  %>% # reorder rows
              left_join(select(datasets, name, short_name),
                        by=c("dataset" = "short_name" )) %>%
              select(-dataset) %>%
              mutate(dataset = as.factor(name),
                     dataset = plyr::revalue(dataset,
                     c("Infant directed speech preference" ="IDS preference",
                     "Statistical sound category learning"= "Statistical sound learning",
                     "Label advantage in concept learning" =  "Concept-label advantage"))) %>%
              mutate(egg_string.f = sub("(0)", "(< .01)", egg_string.f, fixed = T),
                     egg_string.all = sub("(0)", "(< .01)", egg_string.all, fixed = T)) %>%
  select(dataset, egg_string.all, egg_string.f) %>%
  kable(col.names = c("Phenomenon", "funnel skew (all conditions)",
                      "funnel skew (excluding outliers)"),
        align = c("l", "r", "r"))
```

Phenomenon	funnel skew (all conditions)	funnel skew (excluding outliers)
IDS preference	1.5 (0.13)	0.46 (0.65)
Phonotactic learning	-1.43 (0.15)	-1.43 (0.15)
Vowel discrimination (native)	8.55 (< .01)	8.18 (< .01)
Vowel discrimination (non-native)	3.86 (< .01)	3.24 (< .01)
Statistical sound learning	-2.99 (< .01)	-1.89 (0.06)
Word segmentation	2.59 (0.01)	2.8 (0.01)
Mutual exclusivity	8.26 (< .01)	5.2 (< .01)
Sound symbolism	1.42 (0.16)	1.42 (0.16)
Concept-label advantage	1.37 (0.17)	1.37 (0.17)

Phenomenon	funnel skew (all conditions)	funnel skew (excluding outliers)
Online word recognition	2.61 (0.01)	2.61 (0.01)
Gaze following	3.3 (< .01)	3.3 (< .01)
Pointing and vocabulary	1.25 (0.21)	1.25 (0.21)

P-curves

When available, we calculated p-values based on test statistics reported in the paper. However, when unavailable, we calculated p-values based on raw descriptive statistics (means and standard deviations) or reported effect sizes (the method used for IDS Preference). The main text shows the results of the p-curve analysis based on p-values derived by both approaches. Here, we compare these results to the same analysis on the subset of p-values derived only from reported test statistics. Presented below are p-curves calculated from this subset.

```
ALPHA = 0.05
P_INCREMENT = 0.01

pc.data <- get_all_pc_data(all_data, ALPHA, P_INCREMENT,
    transform = FALSE)

p.source = pc.data %>% select(f.transform, f.value,
    dataset, study_ID, p_round) %>% group_by(dataset) %>%
    summarise(n.total = n(), n.transform = length(which(!is.na(f.transform))),
        sig.p = length(which(p_round < ALPHA))) %>%
    mutate(dataset = plyr::revalue(dataset, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
        `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)")),
        dataset = as.factor(dataset), dataset = gdata::reorder.factor(dataset,
            new.order = c(2, 6, 10, 11, 9, 12, 4, 8,
                3, 5, 1, 7))) %>% mutate(stat_only = ifelse(n.total >
    n.transform, 1, 0)) %>% arrange(-stat_only) %>%
    mutate(prop.ts = 1 - n.transform/n.total, prop.ts.string = as.character(round(prop.ts,
        digits = 2))) %>% as.data.frame()

get.all.CIS.multi <- function(df) {
    ps <- seq(P_INCREMENT, ALPHA, P_INCREMENT)
    props = ps %>% map(function(p, d) {
        sum(d == p)
    }, df$p_round) %>% unlist()
    cis = MultinomialCI::multinomialCI(props, alpha = ALPHA)
    data.frame(dataset = df$dataset[1], p = ps, ci.lower = cis[,
        1], ci.upper = cis[, 2])
}

ci.data = pc.data %>% split(.$dataset) %>% map(function(data) get.all.CIS.multi(data)) %>%
    bind_rows() %>% mutate(dataset = as.factor(dataset),
    dataset = plyr::revalue(dataset, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
```

```
        `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)")))

ci.data[ci.data$dataset == "Sound symbolism" & ci.data$p ==
    0.01, "ci.lower"] = 0  # there's only one datapoint for this dataset

pc.data %>% group_by(dataset) %>% do(get_p_curve_df(.,
    ALPHA, P_INCREMENT)) %>% ungroup() %>% mutate(dataset = as.factor(dataset),
    dataset = plyr::revalue(dataset, c(`Statistical sound category learning` = "Statistical sound learni
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
        `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)")),
    dataset = gdata::reorder.factor(dataset, new.order = c(3,
        7, 8, 6, 9, 2, 5, 1, 4))) %>% ggplot() + facet_wrap(~dataset,
    nrow = 3) + geom_ribbon(aes(ymin = ci.lower, ymax = ci.upper,
    x = p), fill = "grey87", data = ci.data) + geom_line(size = 1,
    aes(x = p, y = value, linetype = measure, color = measure)) +
    scale_colour_manual(name = "", values = c("red",
        "green", "blue"), labels = c("Null of no effect",
        "Null of 33% power", "Observed")) + scale_linetype_manual(values = c("dashed",
    "dashed", "solid"), guide = FALSE) + ylab("Proportion p-values\n") +
    xlab("p-value") + geom_text(aes(label = paste("prop. test stat. = ",
    prop.ts.string, "\nnum. sig. ps = ", sig.p), x = 0.028,
    y = 0.8), data = p.source, colour = "black", size = 2,
    hjust = 0) + theme_bw() + theme(legend.position = "top",
    legend.key = element_blank(), legend.background = element_rect(fill = "transparent"),
    strip.text.x = element_text(size = 9), axis.title = element_text(colour = "black",
        size = 12), panel.margin = unit(0.65, "lines"),
    strip.background = element_rect(fill = "grey"))
```

We next compare the test of right-skew presented in the main text for both the full set of p-values and those only derived from test statistics. In no case does the significance of the test differ between the two analyses.

```
stouffer.data = pc.data %>% group_by(dataset) %>% do(data.frame(stouffer = stouffer_test(.,
    ALPHA))) %>% filter(stouffer.pp.measure == "ppr.full") %>%
    full_join(datasets %>% select(name, short_name),
        by = c(dataset = "name")) %>% select(short_name,
    stouffer.Z.pp, stouffer.p.Z.pp) %>% mutate_each_(funs(round(.,
    digits = 2)), vars = c("stouffer.p.Z.pp", "stouffer.Z.pp")) %>%
    mutate(stouff_string = ifelse(is.na(as.character(stouffer.Z.pp)),
        "", paste0(stouffer.Z.pp, " (", stouffer.p.Z.pp,
            ")"))) %>% mutate(stouff_string = sub("(0)",
    "(< .01)", stouff_string, fixed = T)) %>% select(dataset,
    stouff_string)

stouffer.data_all = get_all_pc_data(all_data, ALPHA,
    P_INCREMENT, transform = TRUE) %>% group_by(dataset) %>%
    do(data.frame(stouffer = stouffer_test(., ALPHA))) %>%
    filter(stouffer.pp.measure == "ppr.full") %>% full_join(datasets %>%
    select(name, short_name), by = c(dataset = "name")) %>%
    select(short_name, stouffer.Z.pp, stouffer.p.Z.pp) %>%
    mutate_each_(funs(round(., digits = 2)), vars = c("stouffer.p.Z.pp",
        "stouffer.Z.pp")) %>% mutate(stouff_string = ifelse(is.na(as.character(stouffer.Z.pp)),
    "", paste0(stouffer.Z.pp, " (", stouffer.p.Z.pp,
```

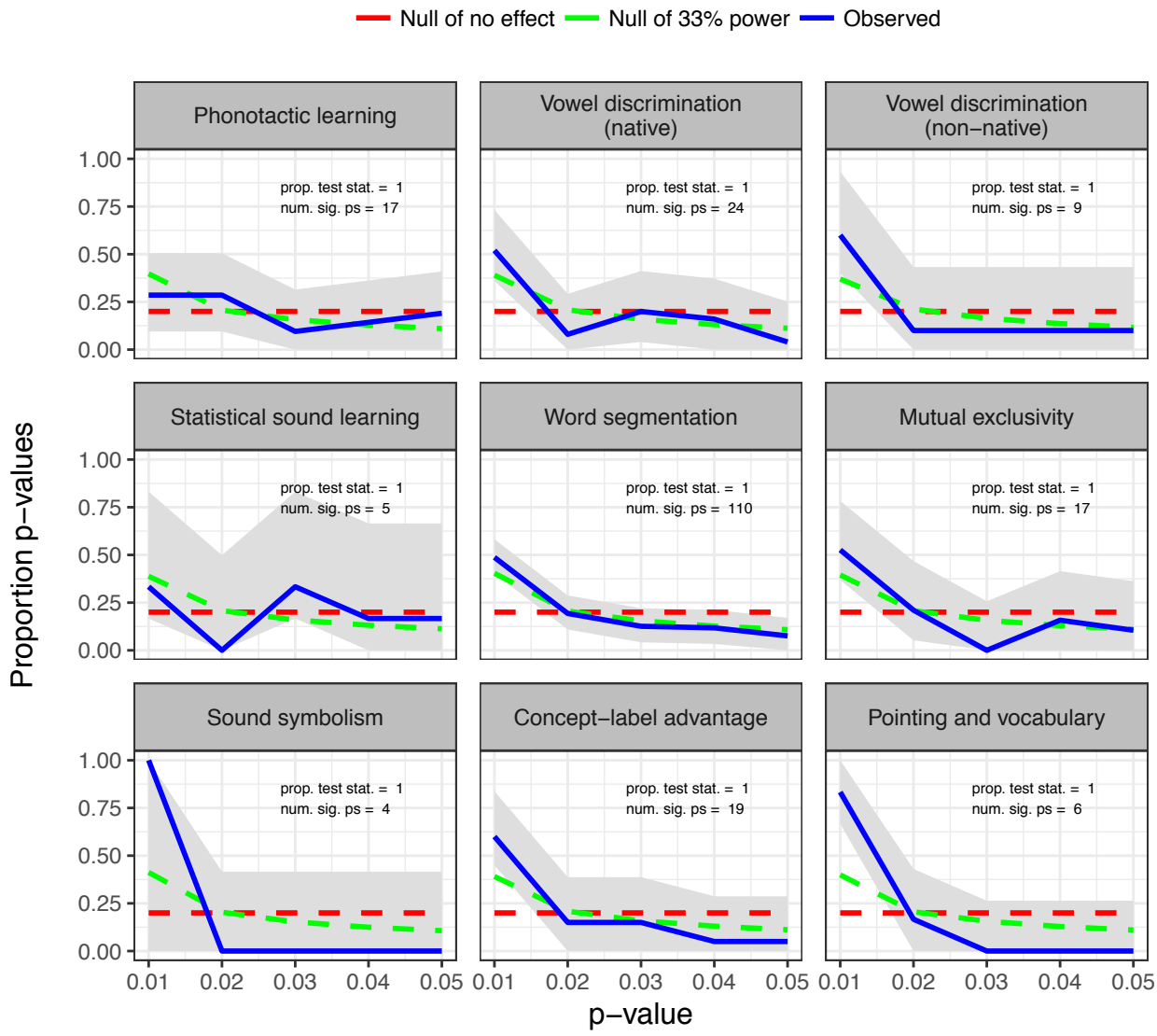


Figure S2: In the main text, we calculate p-curves based on all conditions in the dataset. In cases where a p-value was not directly available from the reported test statistic, we calculated a p-value based on a significance test using the reported means and standard deviations. The table compares the test of right-skew (Stouffer method) for this full dataset, as reported in the main text, to the subset of conditions for which p-values were directly available. Error bars are 95% confidence intervals calculated from a multinomial distribution.

```
        ")"))) %>% mutate(stouff_string = sub("(0)",
    "(< .01)", stouff_string, fixed = T)) %>% select(dataset,
    stouff_string)

# p-curve data using from all conditions (same as
# reported in paper)
pc.data.all <- get_all_pc_data(all_data, ALPHA, P_INCREMENT,
    transform = TRUE)

left_join(stouffer.data_all, stouffer.data, by = "dataset") %>%
    .[c(2, 6, 10, 11, 9, 12, 4, 8, 3, 5, 1, 7), ] %>%

kable(col.names = c("Phenomenon", "p-curve skew  (all conditions)",
    "p-curve skew (p-values only from test-statistics)"),
    align = c("l", "r", "r"))
```

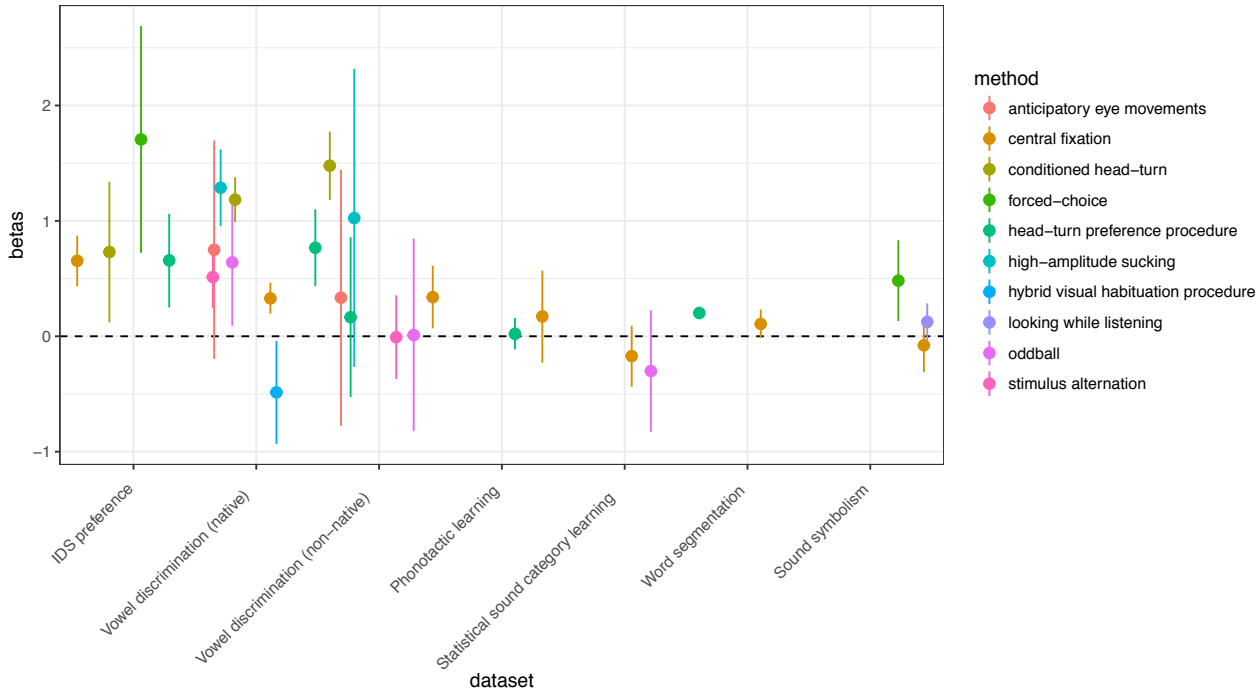
Phenomenon	p-curve skew (all conditions)	p-curve skew (p-values only from test-statistics)
Infant directed speech preference	-10.4 (< .01)	
Phonotactic learning	-1.52 (0.06)	-1.52 (0.06)
Vowel discrimination (native)	-9.76 (< .01)	-5.14 (< .01)
Vowel discrimination (non-native)	-8.89 (< .01)	-3.24 (< .01)
Statistical sound category learning	-1.03 (0.15)	-0.65 (0.26)
Word segmentation	-9.4 (< .01)	-9.82 (< .01)
Mutual exclusivity	-12.87 (< .01)	-5 (< .01)
Sound symbolism	-5.56 (< .01)	-5.1 (< .01)
Label advantage in concept learning	-4.79 (< .01)	-4.54 (< .01)
Online word recognition	-14.51 (< .01)	
Gaze following	-18.66 (< .01)	
Pointing and vocabulary	-6.33 (< .01)	-6.33 (< .01)

Method heterogeneity

The plot below presents model coefficients for method for datasets with more than one method. Coefficients are estimated from random-effects meta-analytic models. For the most part, we see that method only has a small influence on the effect size within a given phenomenon. There are exceptions, however: For example, for Sound Symbolism the forced choice method has an overall larger effect size than other methods.

```
single_method_datasets = all_data %>% group_by(dataset) %>%
    summarise(n_methods = length(levels(as.factor(method)))) %>%
    filter(n_methods == 1) %>% .[["dataset"]]

method.betas = data.frame()
for (i in 1:length(datasets$name)) {

    if (!(datasets$name[i] %in% single_method_datasets)) {
        d = filter(all_data, dataset == datasets$name[i])
        model = metafor::rma(d_calc ~ method - 1, vi = d_var_calc,
            data = d, method = "REML")

        d = data.frame(dataset = datasets$name[i],
            method = row.names(model$b), betas = model$b,
```

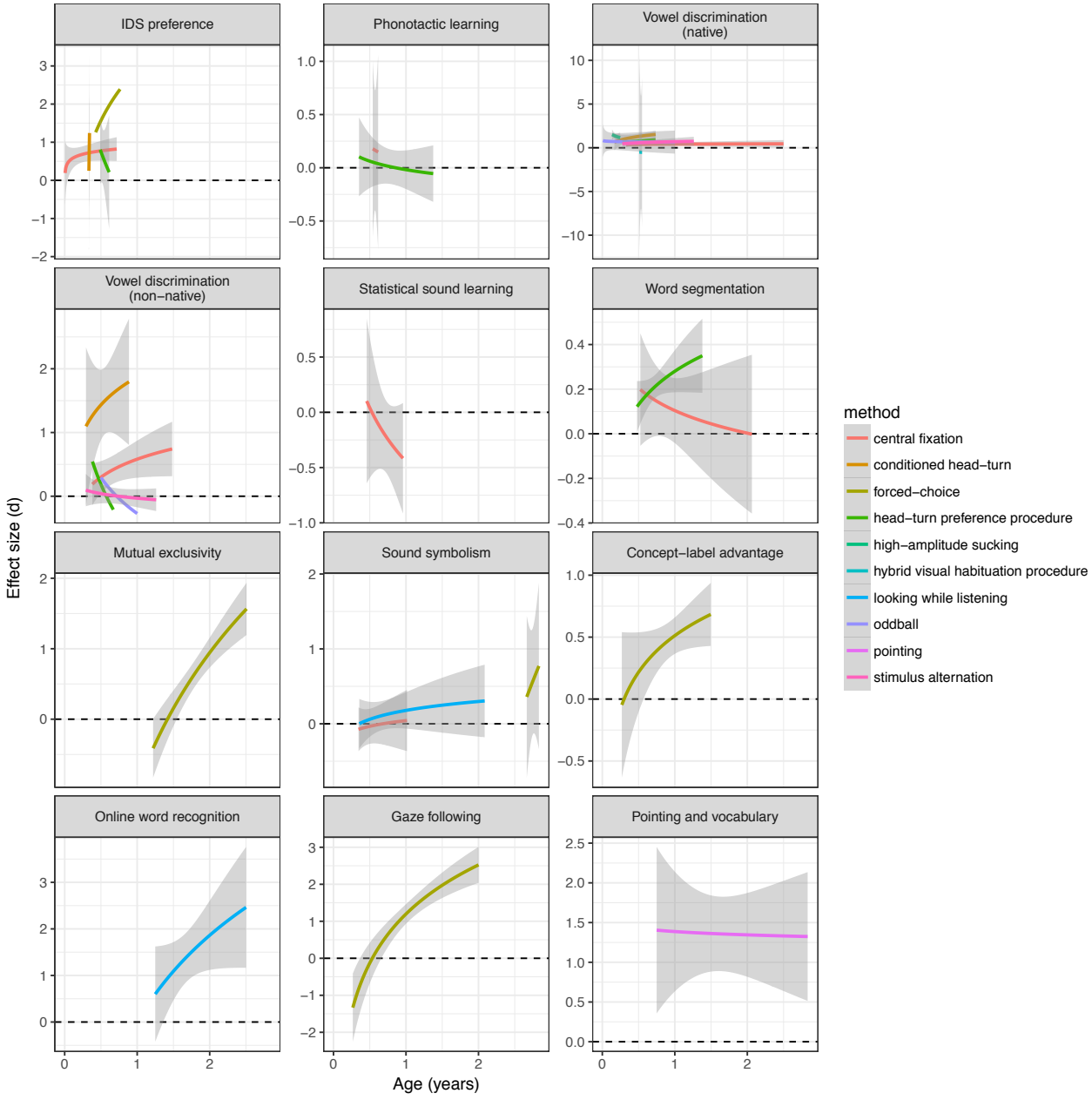
```
            ci.lb = model$ci.lb, ci.ub = model$ci.ub,
            row.names = NULL)

        method.betas = rbind(method.betas, d)
    }
}

method.betas = method.betas %>% mutate(dataset = as.factor(dataset),
    dataset = plyr::revalue(dataset, c(`Infant directed speech preference` = "IDS preference")),
    method = gsub("method", "", method))

ggplot(method.betas, aes(x = dataset, y = betas, ymin = ci.lb,
    ymax = ci.ub, color = method)) + geom_hline(aes(yintercept = 0),
    linetype = "dashed") + geom_pointrange(position = position_jitter(0.5)) +
    theme_bw() + theme(axis.text.x = element_text(angle = 45,
    hjust = 1.1))
```



The plot below presents the developmental trajectory of each phenomenon, with a separate color for each method. Lines show log-linear model fits. Word Segmentation shows the most notable interaction between age and method: Effect sizes increase with age for head-turn preference procedure, but decrease for central fixation.

```
all_data %>% filter(mean_age_1/365 < 3) %>% mutate(dataset = as.factor(dataset),
    dataset = gdata::reorder.factor(dataset, new.order = c(2,
        6, 10, 11, 9, 12, 4, 8, 3, 5, 1, 7)), dataset = plyr::revalue(dataset,
        c(`Infant directed speech preference` = "IDS preference",
            `Statistical sound category learning` = "Statistical sound learning",
            `Label advantage in concept learning` = "Concept-label advantage",
            `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
            `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)"))) %>%
    ggplot(aes(x = mean_age_1/365, y = d_calc, color = method)) +
```

```
    geom_hline(yintercept = 0, linetype = "dashed",
        color = "black") + facet_wrap(~dataset, scales = "free_y",
    ncol = 3) + geom_smooth(method = "lm", formula = y ~
    log(x)) + xlab("Age (years)") + ylab("Effect size (d)") +
    theme_bw() + theme(legend.position = "right", legend.key = element_blank(),
    legend.background = element_rect(fill = "transparent"))
```



References

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

Colonnesi, C., Stamsa, G. J., Kostera, I., & Noomb, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, 30, 352–366.

Cristia, A. (in prep.). Infants' phonology learning in the lab.

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

Frank, M. C., Lewis, M., & MacDonald, K. (2016). A performance model for early wordlearning. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis.* Academic Press.

Lammertink, I., Fort, M., Peperkamp, S., Fikkert, P., Guevara-Rukoz, A., & Tsuji, S. (2016). SymBuki: A meta-analysis on the sound-symbolic bouba-kiki effect in infants and toddlers. Poster presented at the XXI Biennial International Congress of Infant Studies, New Orleans, USA.

Lewis, M. & Frank, M. (in prep.). Mutual exclusivity: A meta-analysis.

Lewis, M. & Long, B. (unpublished). A meta-analysis of the concept-label advantage.

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48. URL: http://www.jstatsoft.org/v36/i03/