MetaLab: A platform for cumulative meta-meta-analyses

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

[1] Ecole Normale Supérieure, PSL Research University, Département d'Etudes Cognitives, Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)

[2] University of Pennsylvania, Department of Psychology

[3] Ecole Normale Supérieure, PSL Research University, Département d'Etudes Cognitives, Neuropsychologie Interventionnelle (ENS, EHESS, CNRS)

[4] Stanford University, Department of Psychology, Language and Cognition Lab

[5] Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences

Author Note

Correspondence concerning this article should be addressed to Christina Bergmann, Ecole Normale Supérieure, Laboratoire de Sciences Cognitives et Psycholinguistique, 29, rue d'Ulm, 75005 Paris, France.. E-mail: chbergma@gmail.com

MetaLab: A platform for cumulative meta-meta-analyses

Todo list

- Collapse over non-independent rows (?)
- Add table with different DVs / analyses from eye tracking studies (needs diff coding scheme in metalab? take example MA?)
- visualize power(?) –> is the plot on d vs p enough?
- p-hacking analyses
  - currently issues with p-value calculation for p-curve

Suggestions I am not sure how to implement:

- Show how infants and children are inherently a more "noisy" population

Would exclusion rates do the trick?

Re title comments: I went with the title we submitted in the proposal

Psychology has recently seen a "crisis of confidence", as recent findings challenge both the validity of key findings (Klein et al., 2014) as well as the general replicability rate of published findings CITE-MANYLABS2. In this process, some practices have been discussed as introducing bias and error, starting at data collection, over analysis to publication, (Ioannidis, 2005),SCIUTOPIA1,2. In other words, psychology is today facing the same issues that caused substantial changes in research practices within the medical sciences a decade ago (De Angelis et al., 2004). The present paper discusses to what extent these issues are present in developmental psychology, with a focus on early studies of language comprehension as one example of a largely consistent subfield of child development research.

Relevance of the confidence crisis for studies on child development

The problems underlying recent confidence crises are thought to be true of empirical sciences at large, given the current reward structure CITE-UTOPIA. Specifically, at present researchers are valued on the basis of the quantity and (some measure of) impact of their

publications, and publication in a high-impact venue is dependent, among other factors, on (a) the topic being "hot"; (b) the result being surprising; and (c) the result being statistically significant. One of the obvious consequences of this reward structure is that replications and even conceptual extensions are less likely to be undertaken (since they are not rewarded as much), and if they are, they are unlikely to be published, particularly if they reveal a non-significant result. Furthermore, modeling suggests that these issues may be exacerbated depending on the characteristics of a given subfield (Ioannidis, 2005), specifically in fields where both the underlying effect sizes and typical sample sizes are small.

All of these descriptions are highly relevant to developmental studies, particularly those focusing early childhood. Since the population under investigation is costly to recruit and difficult test, there is a pressure towards small sample sizes. Moreover, the sample is typically less consistent, as children differ a great deal in their individual developmental trajectory [CITATION?], leading to larger variance and consequently smaller underlying effect sizes. At the confluence of these two factors, one expects to find habitually underpowered studies, which is problematic both when assessing whether an effect is truly present or absent, and when estimating its magnitude for theory building and study planning.

There is a conceptually separate set of issues, which is nonetheless relevant given the aforementioned warning signs for developmental psychology. One of the habits that has become under attack recently pertains to flexibility in data collection and analysis, as arguably this flexibility exacerbates the incidence of false positives for various reasons (Ioannidis, 2005). [More details here? Avoid redundancy with later section.]

In the next sections, we discuss key issues that might undermine the reliability of child development studies for the reasons mentioned above, followed by a principled assessment of possibly questionable practices in language development research as an example case. The use of a large-scale meta-analytic dataset with unprecedented detail and standardization allows for rich analyses that can diagnose the state of a whole subfield and make principled recommendations for future research. The crucial difference between the present work and

single studies resides in the quantitative assessment of study outcomes, instead of a binary result, namely whether the null hypothesis can be rejected with sufficient confidence or not. To preview our results, we find [xxx fill in when results are final xxx]. We end by describing how our the investigations conducted here can be extended to other topics of developmental research.

Key concepts for evaluating the state of a field

suggestions for a different header welcome

Replication and replicability. Replicability is a core concept in the recent crisis, as exactly this property of scientific studies (and potentially whole sub-fields) has turned out to be surprisingly problematic. We define the concept here, as authors vary in their understanding of what constitutes a replication. Replicating a study means (in the context of this paper) conducting a conceptually similar experiment with new stimuli and in a slightly different population but following the same procedure and analyses (based on the published report), tapping into the same phenomenon, and with the same outcome as previously reported (allowing for a margin of error). Being able to (repeatedly) successfully replicate a study can be taken as an indicator that the phenomenon under investigation is true and therefore that theories can be built on it. In addition, varying populations and using different, yet comparable stimuli, assesses generalizability across presumably irrelevant dimensions. If an effect does not generalize across stimuli or populations, it is possible that a previously unknown limitation has been uncovered, which needs to be specified for future replications and within all theories building on the general effect. It is alternatively possible that the initially reported effect does not exist, in that case most attempts to replicate, especially with appropriate power (as discussed in the next section) given the expected size of an effect, should fail. Replicability can be assessed when aggregating all studies that aim to tap into a given phenomenon and assessing whether – taking all evidence together – the effect is statistically different from 0. A next step consists of comparing the direction and

magnitude of initial reports and replications.

Effect sizes. The classical experimental result is measured in terms of a binary distinction: Was the *p*-value significant or not? Beyond answering that simple question, *p*-values cannot offer a quantification of a phenomenon nor can they directly be compared. However, both properties are desirable, especially in a developmental context. We would like to address questions such as how much children improve as they mature or when an ability emerges. Comparing significant and non-significant results to determine the onset of an ability is not a suitable use of *p*-values. The null hypothesis, for example that two groups do not differ, is in fact not being tested, so it cannot be considered confirmed when observing a high *p*-value. Instead, *p*-values can only support rejections of the null hypothesis with a certainty that the data at hand are incompatible with it below a pre-set threshold.

Effect sizes quantify a phenomenon by standardizing differences between groups (either of observations in within-participant designs or of participants) by the variance of each group.[1] Thereby, statements such as improvements with age become statistically founded. Two counter-intuitive fact about effect sizes should be noted. First, a large effect can coincide with a non-significant *p*-value, for example in the presence of large variance around the mean and conversely a significant *p*-value might indicate a small effect, provided the estimate is sufficiently precise. Second, and somewhat related, numeric differences between group means do not determine effect sizes. Again, the variance around those means plays a crucial role, scaling the effect size accordingly.

Sample size and statistical power. Power refers to the probability of detecting an effect and correctly rejecting the null hypothesis given that it is present in a population. The larger the sample size, the higher power. Due to its dependence on the effect size under investigation, and given that infant and child studies are expected to be more noisy compared to adult experiments, we expect lower effect sizes. In tandem with the constraints

[1] The effect size defined here is one of three types, we focus on standardized mean differences for clarity, but note that other paradigms might require a different effect size family.

placed upon researchers in testing this sensitive population, it can be assumed that underpowered studies of child development are the norm.

In developmental psychology, reasons for sample size decisions are rarely reported, it is thus unclear how (or if) decisions about sample size are made before commencing data collection. Formal prospective power calculations are as yet rare, especially those based on multiple studies. Alternatively, it is possible that resource limitations determine sample size, as recruitment can be difficult and is very costly.

Underpowered studies specifically pose a problem for branches of developmental studies that interpret both significant and non-significant findings; for example when tracking the emergence of an ability as children mature or when examining the boundary conditions of an ability. Even in the most rigorous study design and execution, null results will occur ever so often; for example in a study with 80% power (a number typically deemed sufficient), every fifth result will not reflect that there is a true effect present in the population. Disentangling whether a non-significant finding indicates the absence of a skill, random measurement noise, or the lack of experimental power to detect this skill reliably and with statistical support is impossible based on p-values.

A second problem emerges when underpowered studies yield significant outcomes, as the effects reported in such cases will be over-estimating the true effect. This makes appropriate planning for future research which aims to build on this report more difficult, as sample sizes will be too small, leading to null-results which do not speak to the phenomenon under investigation. This poses a serious hindrance for work building on seminal studies, including replications across languages and extensions. However, aggregating over such null-results using a graded estimate, i.e. a standardized effect size, can reveal whether a phenomenon is present in the population and correct for the initial over-estimation. In short, even a true positive result is insufficient in the quest for the truth when it is underpowered.

To investigate the status quo, we first compute typical power across a range of phenomena in early language acquisition, and explore which effect sizes are detectable with

sample sizes in the included studies. Further, we investigate how researchers might determine sample sizes (for example by following the first paper in a literature), and whether they take into account sensitivity of methods used.

Procedural variability. Improving our procedures can be considered both an economical and ethical necessity, because our population is difficult to recruit and test. For this reason, developmentalists often "tweak" paradigms and develop new ones to increase reliability and robustness, all with the aim of obtaining a clear signal. Especially given the time constraints, we aim for a maximum of data in the short time span infants and children are willing to participate in a study. Emerging technologies, such as eye-tracking and tablets, have been eagerly adopted (M. C. Frank, Sugarman, Horowitz, Lewis, & Yurovsky, 2016). As a result, multiple ways to tap into the same phenomenon have been developed, consider for example the fact that both headturn-based paradigms and the measurement of eye movements have been employed to measure infant-directed speech preference (Dunst, Gorman, & Hamby, 2012), Manybabies1]. It remains an open question to what extent these different methods lead to comparable results. It is possible that some are more robust, but it remains difficult to extract such information based on studies that use different materials and test various age groups. Aggregating over experiment results, however, allows us to extract general patterns of higher or lower noise via comparison of effect sizes, which are directly affected by the variance of the measurement.

We will assess in how far the different methods used to test the same construct vary in their sensitivity. Further, taking possible resource limitations into account, we consider drop-out rates as a potential measure of interest and discuss whether higher exclusion rates coincide with more precise measures, yielding higher effect sizes.

P-hacking. Undisclosed flexibility during data collection and analysis is a problem independent of the availability of various methods to conduct infant studies. During data collection, a number of practices can inflate the number of significant p-values, effectively rendering p-values and the notion of statistical significance meaningless (Ioannidis, 2005).

First, flexible stopping rules, including adding observations when a test statistic is "promising" or stopping data collection when a result is "significant" increases the likelihood to obtain a "significant" outcome. Another form of p-hacking is measuring several dependent variables and conducting multiple significance tests with each variable and with a combination of the variables. In developmental research, this problematic practice encompasses computing several dependent variables (such as mean scores, difference scores, percentages, and so on) based on the same measured data as well as selectively excluding trials and re-testing the new data. Next, multiple conditions that selectively can be dropped from the final report increase the number of significance tests. Finally, it is problematic to post hoc introduce covariates, most prominently gender, and test for an interaction with the main effect. Finally combining two or more of these strategies again inflated the number of significant results. All these practices might seem innocuous and geared towards "helping" an effect to emerge that the researcher believes to be real.

A related issue is that there is little standardization in data analyses practices. The same type of data, such as eye movements, can be assessed with various statistical procedures and little consensus has emerged over time methods have been applied. While the availability of multiple analyses can be tempting to adopt practices that undermine the reliability of results, it it worthwhile to adopt new analyses methods to reduce noise. However, researchers might try out multiple statistical analyses and only report those with a p-value below the significance threshold. From a single report it is not possible to assess whether such analytic flexibility led to an inflation of significant results, but we use tools that are based on cumulative science in the present work to distinguish between optimizing informativeness and unsavory practices.

A "symptom" of such practices is a distribution of p-values with increased frequency just below the significance threshold. P-curves test for this problem, but they come with some limitations and only consider statistically significant reports [p-curve citation, limitation citation].

[Insert example table of different DVs here? I am worried it is too accusatory.]

Publication biases. As mentioned previously, current incentives including publication of data in a prestigious journal, are geared towards surprising and statistically significant studies. However, even when an effect is robust and tested with sufficiently high participant numbers, null results are expected to occur. This becomes even more pressing in a field with small effect sizes and low numbers of participants. For an accurate estimate of the true effect it is crucial to have access to all results, to avoid overestimations. However, we expect that studies remain in the file-drawer and never see the light of day. One reason is that a failure to obtain an expected significant result is often ascribed to the researcher's skill or an unknown flaw in the experiment [citation?].

We will investigate publication biases in our data with standard meta-analytic tool and further discuss how meta-analyses in general, and repositories such as MetaLab in particular, can be a "home" for null results. A cumulative view can help isolate factors that systematically lead to effect sizes closer to zero, or what is often called boundary conditions. Prerequisite for such analyses is a systematic and consistent reporting of such factors; if they are relevant this should however be the case according to standard scientific practice.

Methods

Source data: MetaLab

In this paper, we extract measures of interest from meta-analyses of child language development. Meta-analyses are built on a collection of standardized effect sizes on a single, well-defined phenomenon. By accumulating effect sizes and weighting them by their reliability (effectively the sample size), it is possible to compute an estimate of the population effect, as well as its structured variance. By harnessing data from hundreds of studies, we can quantify patterns important for experimental practices. Furthermore, combining multiple meta-analysis – each centered on a different research question – allows us to assess whether current practices differ across different topics.

Given that all 11 meta-analyses we discuss in this paper focus on language acquisition in early childhood, our suggestions will be most relevant to this sub-field. We present our methods and results to researchers on developmental psychology in general to encourage others to build similar meta-meta-analyses, thus allowing them to explore the state of their own sub-fields and to improve their practices if necessary. The analyses in this paper are based on MetaLab, an online collection of meta-analyses on early language development. Currently, MetaLab contains 11 meta-analyses, but it is open to submissions and updates. The present analyses thus are a snapshot; through dynamic reports on the website, and by downloading the freely available data, it is continuously possible to obtain the most recent results.

In MetaLab, parts of each meta-analysis are standardized to allow for the computation of common effect size estimates and for analyses that span across different phenomena. These standardized variables include study descriptors (such as citation and peer review status), participant characteristics (including mean age, native language), methodological information (for example what dependent variable was measured), and information necessary to compute effect sizes (number of participants, if available means and standard deviations of the dependent measure, otherwise test statistics, such as t-values or F scores).

MetaLab contains datasets that address phenomena ranging from infant-directed speech preference to mutual exclusivity, sampled opportunistically based on data collected with involvement of (some) authors of this paper (n=9 datasets) or they were extracted from previously conducted meta-analyses related to language development (n=2, i.e. (Colonnesi, Stams, Koster, & Noom, 2010, Dunst et al. (2012))). In the former case, we attempted to document as much detail as possible for each entered experiment (note that a paper can contain many experiments). Detailed descriptions of all phenomena covered by MetaLab, including which papers and other sources have been considered, can be found on the companion website at http://metalab.stanford.edu and in the supporting information.

Further, a throughout investigation into data quality within MetaLab and a

meta-meta-analyses have been conducted based on the same data (M. Lewis et al., 2016). Meta-analyses do not rely on one (possibly inaccurate) study outcome, be it significant or not. Despite their overall utility, meta-analyses are not frequently conducted in most branches of developmental psychology. Instead, narrative summaries are the dominant tool to build theories, and that single studies are cited as evidence for the presence or absence of an ability instead of meta-analyses.

Statistical approach

As dependent measure, we report Cohen's *d*, a standardized effect size based on comparing sample means and their variance. This effect size was calculated when possible from means and standard deviations across designs with the appropriate formula. When these data were not available, we used test statistics, more precisely t-values or F scores of the test assessing the main hypothesis. We also computed effect size variance, which allows to weigh each effect size when aggregating across studies. The variance is mainly determined by the number of participants; intuitively effect sizes based on larger samples will be weighted higher. Note that for research designs testing participants in two conditions that need to be compared (for example exposing the same infants to infant- and adult-directed speech), correlations between those two measures are needed to estimate the effect size variance. This measure is usually not reported, despite being necessary for effect size calculation. Some correlations could be obtained through direct contact with the original authors (see e.g., (C. Bergmann and Cristia, 2016) for details), for others we estimated this factor based on the information in our database.

Meta-analytic model. To aggregate effect sizes within a phenomenon, we used a multilevel approach, which takes into account not only the effect sizes and their variance of single studies, but also that effect sizes from the same paper will be based on more similar studies than effect sizes from different papers (Konstantopoulos, 2011), implemented in the metafor package (Viechtbauer, 2010) of R (R Core Team, 2016). We excluded as outliers

effect sizes that were more than three standard deviations away from the median effect size within each dataset, thus accounting for the difference in median effect size across phenomena.

P-curves. For analyses involving *p*-values, we re-computed *p*-values from our effect-size estimates. This is due to two main reasons: First, we did not have the same information available for all data points, even within the same meta-analysis. For example, in XXX, XXX% of the effect sizes were calculated based on t-values, XXX% based on group-level means and standard deviations, and XXX% based on F-scores. In addition, some datasets only contain effect sizes, because they are based on extant meta-analyses. Second, *p*-values are not always computed and reported correctly or consistently (Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2016). To ensure a consistent relationship between *p*-values and effect sizes, we thus opted for recalculation. The recalculation pipeline is as follows: We transform Cohen's *d* into Pearson's *r*, from which it is possible to calculate a *t*-value.

P-curves are based on the subset of significant *p*-values (Simonsohn, Nelson, & Simmons, 2014). In the absence of practices to cross the threshold between non-significant and significant results, a right-skewed distribution is to be expected – assuming the effect under investigation is truly present in the population. If a non-existent effect is being measured, all *p*-values are equally likely to occur, and the curve is expected to be flat. A third option is a left-skew, with an increase of observed *p*-values just below the significance threshold. This indicates severe p-hacking, including adding and removing samples and/or predictors, and conducting multiple statistical analyses (Ioannidis, 2005). The three possible outcomes are compared to the observed p-curves for each dataset.

Assessing publication bias. There are numerous ways to estimate whether the published literature is biased. The most common and straightforward is an assessment of funnel plot asymmetry. A funnel plot displays effect sizes against their variance (with 0 being plotted up). The expectation in the absence of biases is that effect sizes are equally distributed around the meta-analytic mean, and that the are spread out more the larger their

variance, creating a triangle-like shape. Biases can lead to distortions in this distribution. The large the asymmetry, the more likely a bias is. We quantify funnel plot asymmetry with a rank correlation test implemented in the metafor package (Viechtbauer, 2010).

A second analysis pertains to the relationship between observed effect sizes in single studies and the associated sample size. The smaller the effect size, the larger the sample needed for a significant p-value. If sample size decisions are made before data collection and all results are published, we expect no relation between observed effect size and sample size. A significant non-parametic correlation indicates that only those studies with significant outcomes were published (Begg & Mazumdar, 1994).

We would like to point to limitations of this analysis in our diverse dataset. It is not appropriate for meta-analyses combining different effects to draw strong conclusions from a correlations. Consider for example word segmentation, where researchers might add additional factors, such as speaker identity, to further investigate infants' emerging abilities. If the expected interaction effect is smaller, investigators could opt for larger sample sizes.

Results

Measures of a typical study on early language acquisition

Table 1 provides a summary of typical sample sizes and effect sizes by phenomenon, but before discussing those descriptors in detail, we begin by characterizing the overall snapshot provided by our data. Overall, we observe small sample sizes (the overall median in our dataset is 18). With such a sample size, and assuming a paired t-test based on within-participant comparisons (the most frequent experiment design and test statistic) it is possible to detect an effect in 80% of all studies when Cohen's $d = 0.70$, in other words when investigating a medium to large effect. When comparing groups, this number increases to Cohen's $d = 0.96$, a large effect.

The above observation concerning sample sizes and which effect size could be detected in a typical study on early language acquisition is in stark contrast with the effect sizes we

actually observe, which tend to fall into ranges of small to medium effects. Taking a closer look at single phenomena, which we characterize along a number of dimensions, such as typical age and the number of studies (and papers) we base our observations on. Each effect size was calculated with a hierarchical random effects model (Viechtbauer, 2010), taking into account increased similarity between studies in the same paper and weighting effect sizes by their variance (driven by the number of participants). Based on the meta-analytical effect size and the median number of participants, we calculated typical power (using the pwr package (Champely, 2015)). We remind the reader that recommendations are for this value to be above 80%, which refers to a likelihood that 4 out of 5 studies show a significant outcome for an effect truly present in the population. It turns out that by and large, studies are underpowered.

Phenomena in MetaLab differ in the age groups typically tested and the age range covered, with the mean age ranging between 4.5 months (infant directed speech preference) and 2.5 years (mutual exclusivity). One might expect a relationship between effect sizes and infant age both for theoretical and practical reasons. On one hand, younger infants might show a smaller effect in general because they are not yet as proficient in their native language, having had less experience, and because they are a more immature in terms of their information processing abilities [CITE]. On the practical side, methods – a topic we will investigate in depth in the next section – might be more noisy for younger infants and they could be a more difficult population to recruit.

While there is no strict linear relationship between infant age and sample size, effect size, and the derived power, we observe a difference between studies typically testing infants younger than one year and those testing older infants. First, sample sizes are much lower for younger infants, which do usually not test more than 20 infants (although all datasets contain studies with larger samples). This is not the case for older children. The only exception is the dataset addressing mutual exclusivity, which habitually tests around 16 children, which is somewhat off-set by a comparatively large effect size. Additionally, the

number of participants tested within each dataset ranges a great deal, between single-digit numbers and in some cases more the tenfold amount. This might indicate that researchers are mostly limited by their resources and participant availability in planning their studies.

Turning to effect size, we see a similar split by age group in our data. Younger infants show both a greater range and include lower effect sizes which fall into the classical range of small effects (Cohen's d below .5), which is not the case for older children. Power is directly related to sample size and effect size, so it is not surprising that typical power is greater for older children. Interestingly, however, there seems to be little to no relationship between effect sizes and number of participants typically tested. For phenomena with large effects, this means that studies are very high-powered (see gaze following, online word recognition, as two examples). For younger children, because sample sizes and effect sizes are both small, power is habitually very low, and the only dataset which typically achieves appropriate power near 80% is non-native vowel discrimination. For older children, power is solely caused by lower effect sizes. The lack of a relationship between overall meta-analytic power and sample size might indicate that researchers' experiment planning is not impacted by the phenomenon under investigation. Studies might instead be designed and conducted with pragmatic considerations in mind, such as participant availability.

Besides this very general point, we refrain here from strong conclusions based on the above-discussed observations, since the present dataset is not exhaustive and topics typically investigated in younger children are over-represented. However, we sampled in an opportunistic and thus to some degree random fashion and the phenomena covered span very different aspects of language acquisition and linguistic processing.

How does effect size relate to *p*-values?. The key measure in this paper is, as described in the Introduction, an effect size. In contrast, single experiments are often evaluated by their associated p-value, despite the frequent criticisms and well-documented shortcomings of that measure [citations]. Effect sizes are a continuous measure, while p-values are largely used in a binary way,namely to either reject the null hypothesis or fail to

Table 1

(#tab:Descriptive Information)Descriptions of meta-analyses currently in MetaLab.

Topic	Mean Age (Months)	Median Sample Size (Range)	# Effect Siz
Infant directed speech preference	4.34	20 (10, 60)	
Vowel discrimination (native)	6.54	12 (6, 50)	1
Vowel discrimination (non-native)	7.69	16 (8, 30)	
Sound symbolism	7.89	20 (11, 40)	
Statistical sound category learning	8.16	14.75 (5, 35)	
Word segmentation	8.29	20 (4, 64)	2
Phonotactic learning	10.69	18 (8, 40)	
Label advantage in concept learning	12.36	13 (9, 32)	
Gaze following	14.24	23 (12, 63)	
Online word recognition	18.00	25 (16, 95)	
Mutual exclusivity	23.99	16 (8, 72)	

do so. Figure XXX illustrates this binary character with a horizontal line, along with the continuous aspect of *p*-values.

The figure illustrates the intuition that very high effect sizes are related to significant *p*-values and very low effect sizes near zero lead to non-significant *p*-values. However, the majority of effect sizes we observe falls in a range that with sufficient power, in other words a sufficiently large sample, can lead to a significant outcome. Underpowered studies, in contrast, might observe a similar sized effect but fail to reach significance. It should also be noted that 5 of the 12 meta-analytical effect sizes in this paper fall in a range where single studies are not significant, that is below d = 0.61.

Comparing meta-analytic effect size and oldest paper to estimate power. As Table 1 shows, experimenters are habitually not including a sufficient number of participants to observe a given effect, assuming the meta-analytic estimate is accurate. It

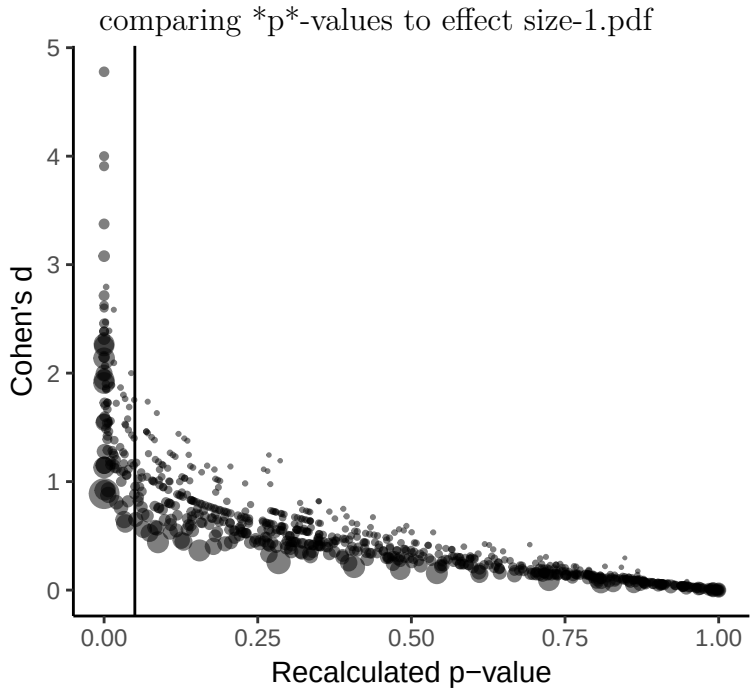

Figure 1. (#fig:Plot comparing *p*-values to effect size)Comparison of a study's effect size and the according *p*-values. Point size reflects sample size. The typical significance threshold of .05 is indicated by a vertical line.

might, however, be possible, that power has been determined based on a seminal paper to be replicated and/or built on. Initial reports tend to overestimate effect sizes (Jennions & Møller, 2002), possibly explaining the lack of power in some datasets. We extracted for each dataset the oldest paper and therein the largest reported effect size and re-calculated power accordingly, using again the median sample size. The results are shown in the table below. It turns out that in some cases, such as native and non-native vowel discrimination, sample size choices match well with the oldest report. The difference in power, noted in the last column, can be substantial, with native vowel discrimination and phonotactic learning being the two most salient examples. Here, sample sizes match well with the oldest report and studies would be appropriately powered if this estimate were representative of the true effect.

To illustrate the disparity between the oldest effect size and the meta-analytic effect, and consequently the difference in power, we plot the difference between both against the

Table 2

(#tab:Compute Meta-analytic d for All Papers)For each meta-analysis, largest d from oldest paper and power, along with the difference between power based on meta-analytic and oldest d.

Meta-analysis (MA)	Oldest Paper	Oldest d	Median Samp
Statistical sound category learning	Maye, Werker, & Gerken (2002)	0.56	
Word segmentation	Jusczyk & Aslin (1995)	0.56	
Mutual exclusivity	Merriman et al. (1989)	0.70	
Label advantage in concept learning	Balaban & Waxman (1997)	0.86	
Vowel discrimination (non-native)	Trehub (1976)	1.02	
Phonotactic learning	Chambers et al. (2003)	0.98	
Sound symbolism	Maurer, Pathman, & Mondloch (2006)	0.95	
Online word recognition	Zangl et al. (2005)	0.89	
Gaze following	Mundy & Gomes (1998)	1.29	
Vowel discrimination (native)	Trehub (1973)	1.87	
Infant directed speech preference	Glenn & Cunningham (1983)	2.39	

oldest effect. This difference is larger as oldest effect size increases, with an average of 0.50 compared with an average effect size of 0.60 (note that we based this on the absolute value). The plot showcases that researchers might want to be wary of large effects, as they are more likely to be non-representative of the true phenomenon compared to smaller initial effects being reported. Especially when making decisions about sample sizes, large effect might thus not be the best guide. Taking the above-mentioned mean values as example, a realistic sample size to ensure 80% power would be 45 participants, instead of 14 participants suggested by the first paper. While these numbers average over research questions and methods, which all influence the specific number of participants necessary, this example showcases that experimenters should take into account as much evidence as available to be able to plan for robust and reproducible studies.

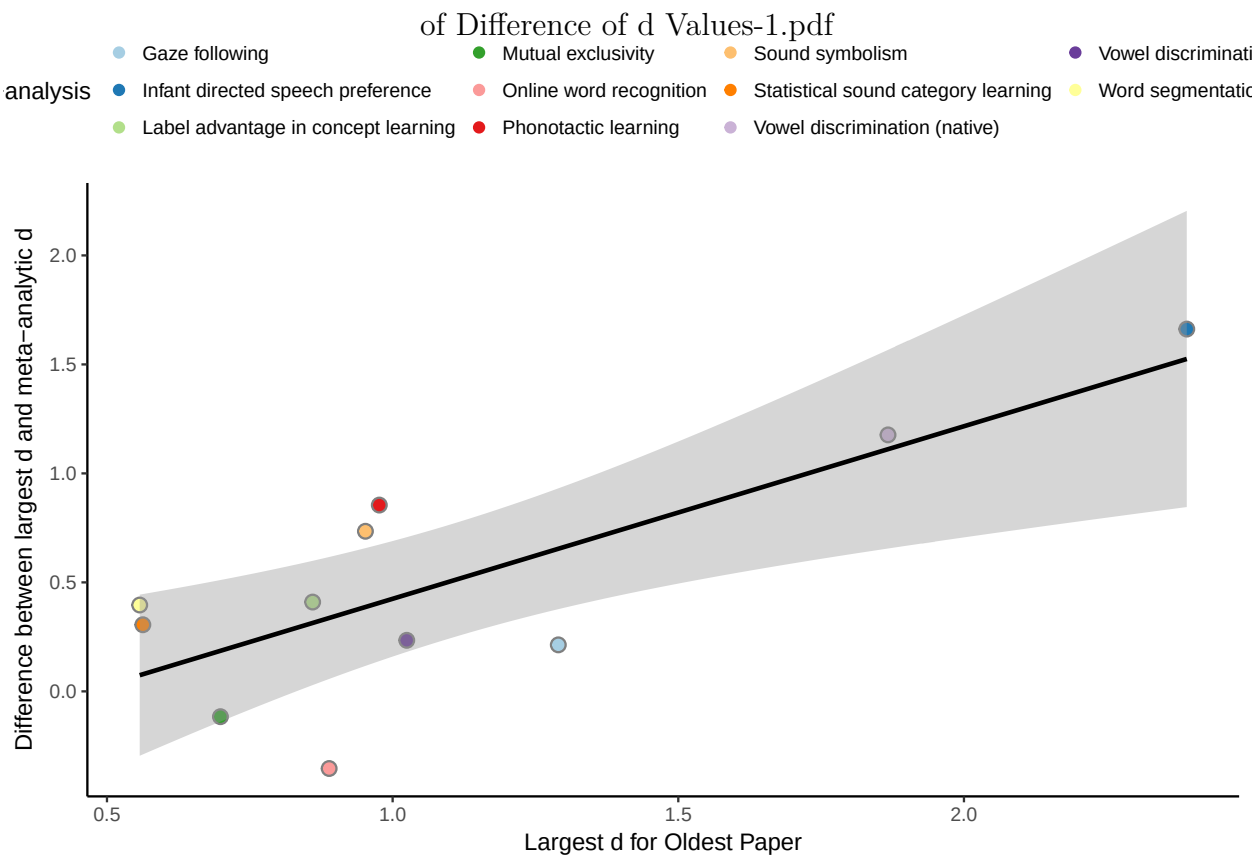

Figure 2. (#fig:Plot of Difference of d Values)Correlation of largest d from oldest paper and difference between oldest d and meta-analytic d.

Procedure comparison

In this section we address how methods might be chosen, adopting two angles. We first take a pragmatic, resource-oriented approach and compare methods with respect to their dropout rate. Then we compare how effect size across phenomena is relating to method choice.

Drop-out rates across methods and age. Choosing a robust method can help increase the power of studies, such that more precise measurements lead to larger effects and thus require fewer participants to be tested. However, the number of participants relates to the final sample and not how many infants had to be invited into the lab. We thus first quantify whether methods differ in their typical drop-out rate, as the available participant

pool might inform method choice. To this end we consider all methods across datasets in MetaLab which have more than 10 associated effect sizes and for which information on the number of dropouts was reported; this information is not always available in the published report, and in the case of the two meta-analyses we added based on published reports, the information was not added. Therefore, the following analyses only cover 6 methods and 224 data points.

The results of the linear mixed effect model predicting dropout rate by method and mean participant age (while accounting for the different effects being tested) are summarized in the table below. The results show that, taking central fixation as baseline, conditioned headturn and stimulus alternation have significantly more drop-outs. Figure XXX underlines this observation, and illustrates the relationship of drop-out rate with age. Overall, stimulus alternation leads to the highest drop-out rates, which lies at around 50% across all age groups. While age is not significantly impacting drop-out rates, it interacts with the different methods. We observe an increase in drop-out rates, which is most prominent in conditioned headturn (a significant interaction) and headturn preference procedure (where the interaction approaches significance).

Interestingly, the methods with lower drop-out rates, namely central fixation and headturn preference procedure, are among the most frequent ones in MetaLab and certainly more frequent than those with higher drop-out rates, indicating that drop-out rate might inform researchers' choices. Being able to retain more participants as a factor in method choice points to the mentioned limitations regarding the participant pool we mentioned before, as more participants will have to be tested to arrive at the same sample size.

The effect of method choice on effect sizes (and thus power). Methods which retain a lot of participants might either be more suitable to test infants, decreasing noise as most participants are on task, or less selective, thus increasing noise as participants who for example are fussy are more likely to enter the data pool. We operationalize precision as the size of the effect measured. Some datasets contain only one method, making it thus

Table 3

(#tab:Method vs excluded)Method vs Dropout

	Estimate	Std. Error	t value
(Intercept)	31.21	4.46	7.00
methodconditioned head-turn	30.62	5.61	5.45
methodforced-choice	-26.42	9.40	-2.81
methodhead-turn preference procedure	-2.33	4.75	-0.49
methodlooking while listening	-6.42	5.37	-1.19
methodstimulus alternation	21.34	4.10	5.21
ageC	0.42	0.44	0.95
methodconditioned head-turn:ageC	2.88	1.16	2.47
methodforced-choice:ageC	-0.22	0.65	-0.33
methodhead-turn preference procedure:ageC	0.96	0.72	1.34
methodlooking while listening:ageC	-0.57	0.80	-0.71
methodstimulus alternation:ageC	-0.26	0.91	-0.29

difficult to disentangle the effect size of a phenomenon with the change of effect size introduced by different methods. To avoid this confound, we limited this investigation to the 4 datasets that contain three or more different methods. We further only investigate those methods that have at least 10 effect sizes in our overall dataset. Thus, the present analyses are limited to 232 observations.

We built a meta-analytic model with the effect size measure Cohen's d as the dependent variable, method and mean age centered as independent variables. The model also includes the variance of d for sampling variance, and paper within meta-analysis as a random effect (because we assume that within a paper experiments and thus effect sizes will be more similar to each other than across papers). Since the model compares one method as the baseline to all other methods, a baseline method had to be chosen. "Central fixation"

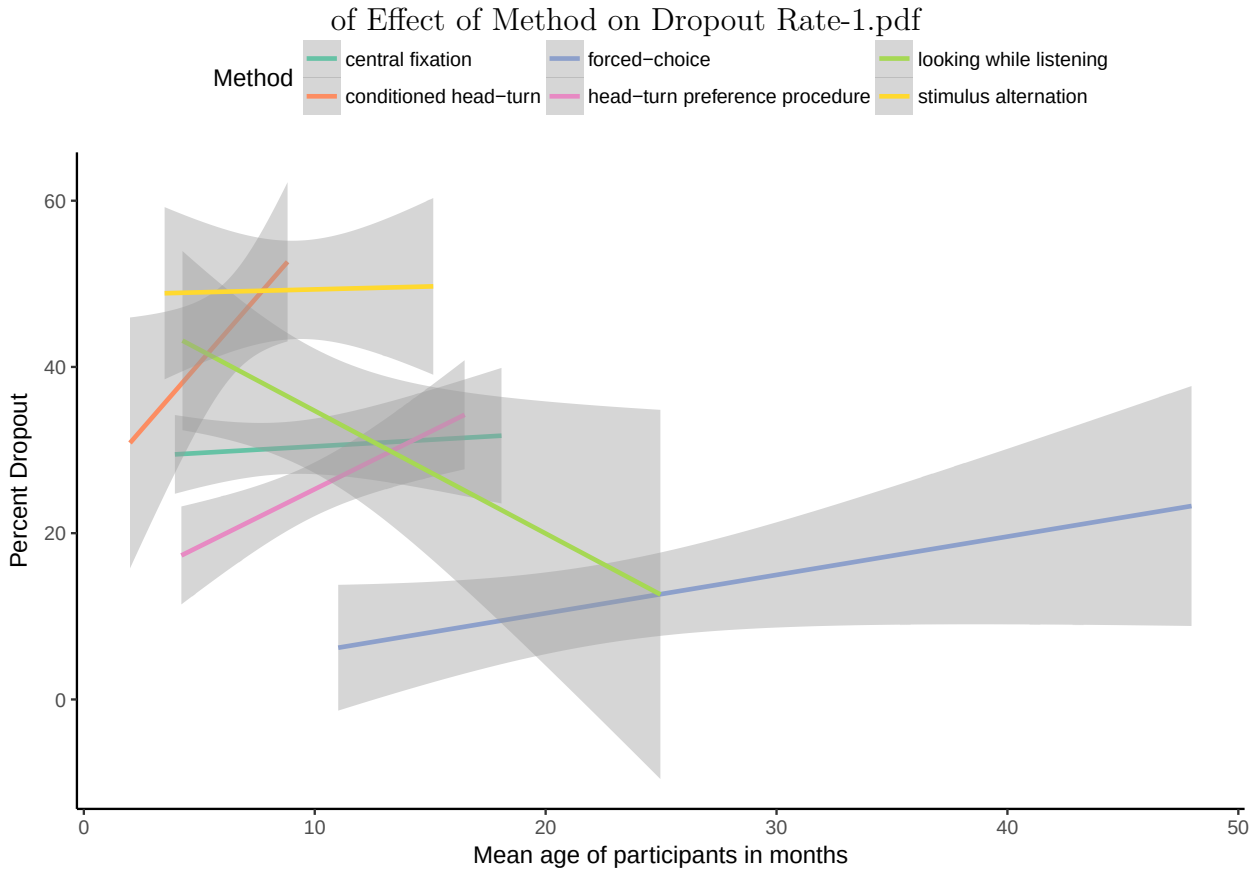

Figure 3. (#fig:Plot of Effect of Method on Dropout Rate)Percent dropout as explained by different methods and mean age of participants.

was included as the baseline method, as it appears most frequently in the 4 datasets included in this analysis (100 times out of 232 total entries of the selected meta-analyses).

The model results in Table XXX show that compared to central fixation only conditioned headturn yields reliably higher effect sizes, all other methods do not statistically differ from this baseline. When factoring in age, no interaction reaches significance, while this factor on its own is marginally below the significance threshold, indicating that as infants mature effect sizes increase across methods – an observation consistent with the view that infants and toddlers become more proficient language users and are increasingly able to react appropriately in the lab.

Comparing our analyses (Table XXX) and Figure YY in this section with those in the

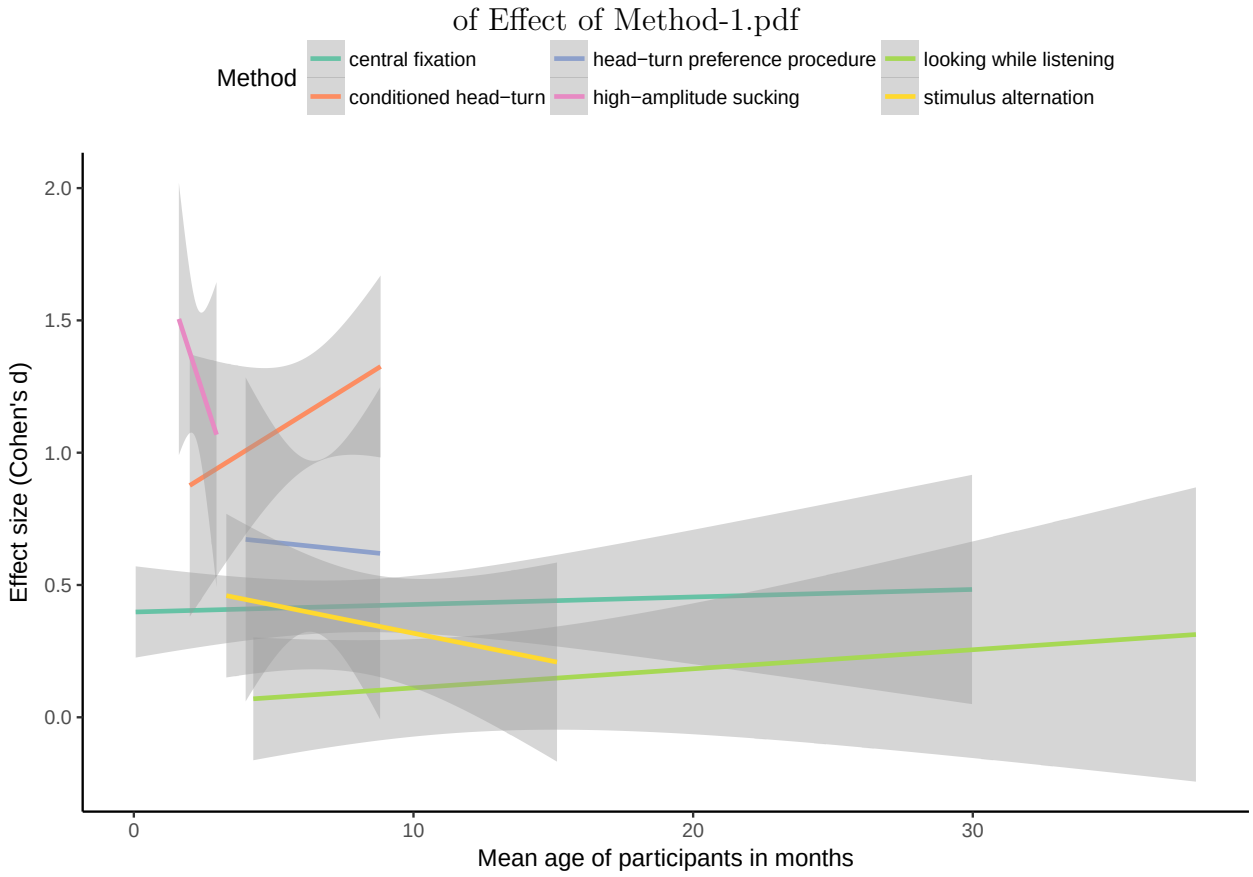

Figure 4. (#fig:Plot of Effect of Method)Effect size as explained by different methods and mean age of participants.

previous section, it seems that high drop-out rates might be offset by high effect sizes in the case of conditioned headturn. While drop-out rates are around 40-50%, effect sizes are above 1. Only high-amplitude sucking seems to generate even higher effect sizes, but for this method we did not have enough information on drop-out rates available, so we cannot examine the relationship between the two. Further, due to the few data points available (13 associated effect sizes) the difference between high-amplitude sucking and central fixation was not significant. Stimulus alternation does not fall into this pattern of high drop-out rates being correlated with high effect sizes, as the observed outcomes are in the range typical for meta-analyses in our dataset.

There is an important caveat to this interpretation that some methods, specifically

conditioned headturn, which have higher dropout rates, are better at generating high effect sizes due to decreased noise (e.g., by excluding infants that are not on task). Studies with fewer participants (thanks to higher drop-out rates) might simply be underpowered, and thus any significant finding is likely to over-estimate the effect. Due to publication biases, we might not have access to all null results using the same method, and the resulting overestimation is directly reflected in our effect size estimate. As a next step, we thus quantify the possible publication biases in our data.

P-hacking

#to be added

Publication biases

Funnel plots, displaying effect sizes of single studies against their variance, show whether observed results are evenly spread around their median. Across datasets, the difference in distributions and range covered in effect sizes is striking, as is the variance in observed precision (points plotted upwards close to zero). The indicated relationship between effect sizes and their variance was assessed with a nonparametric test and turnd out to be significant – indicating publication bias in favor of significant results – for 4 datasets.

Only two datasets turn out to have a significant negative relationship between sample size and effect size, indicating bias. As discussed in the methods section, however, a number of alternative explanations are possible. As soon as researchers are aware that they are measuring a more subtle effect (in this case for example when selecting a contrast that is acoustically more difficult to distinguish) and adjust sample sizes, we expect to observe this negative correlation. However, in both datasets, funnel plot asymmetry was also significant.

Discussion

Concrete recommendations for developmental psychologists

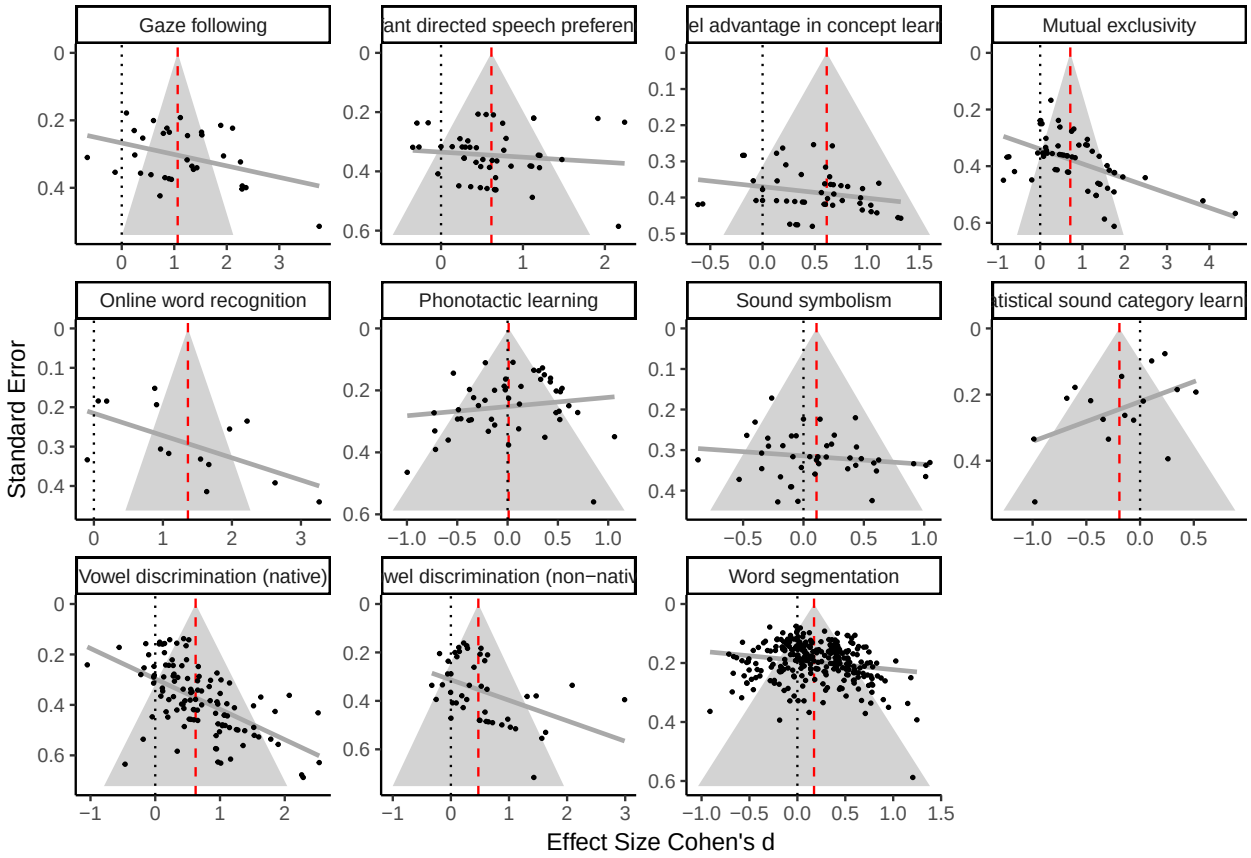

Figure 5. Funnel plots of all datasets with linear regression line indicated in grey. The dashed red line indicates the median effect size, the dotted black line zero. The grey triangle denotes the expected range area of effect sizes in the absence of bias or heterogeneity.

1. **Calculate power prospectively.**

2. **Preregister.**

3. **Increase availability and use of meta-analyses.** To support the improvement current practices, we propose to make meta-analyses available in the form of ready-to-use online tools, dynamic reports, and as raw data. These different levels allow researchers with varying interest and expertise interests to make the best use of the extant record on infant language development, including study planning by choosing robust methods and appropriate sample sizes. There are additional advantages for interpreting single results and for theory building that emerge from our dataset. On one hand, researchers can easily check whether their study result falls within the expected range of

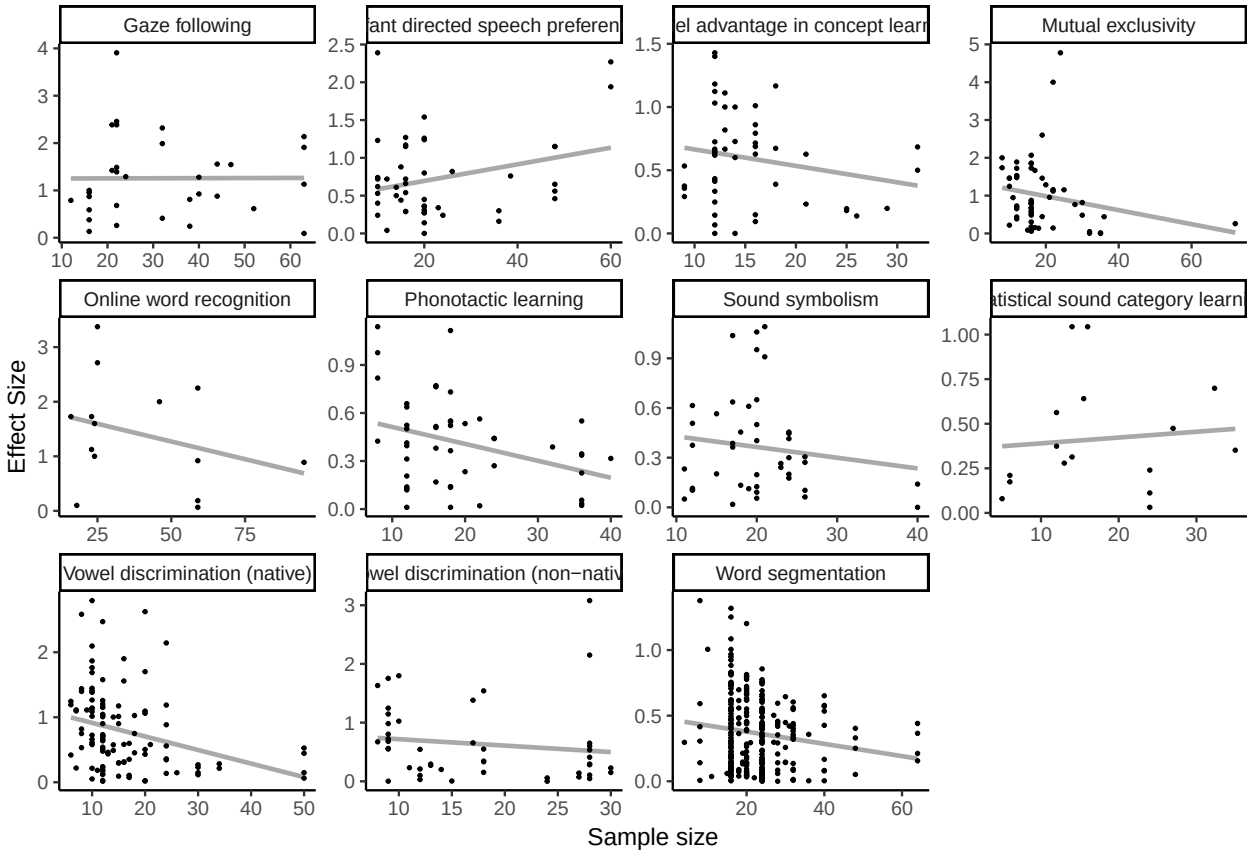

Figure 6. Caption.

outcomes for their research question – indicating whether or not a potential moderator influenced the result. On the other hand, aggregating over many data points allows for the tracing of emerging abilities over time, quantifying their growth, and identifying possible trajectories and dependencies across phenomena (for a demonstration see M. Lewis et al. (2016)). Finally, by making our data and source code open, we also invite contributions and can update our data, be it by adding new results, file-drawer studies, or new datasets. Our implementation of this proposal is freely online available at metalab.stanford.edu.

We have shown that power varies greatly across phenomena and that method choice is important. It turns out, however, that researchers do not choose the most robust methods. This might to be due to a lack of consideration of meta-analytic effect size estimates. One of the reasons for this is a lack of information on and experience in how to interpret effect size estimates and use them for study planning (Mills-Smith, Spangler, Panneton, & Fritz, 2015).

Meta-analyses on infant language development are also rare, as showcased by the fact that the present dataset relied on the authors' involvement, and only two out of XXX meta-analyses used could be extracted from the extant work, and an extensive search in the present literature did not yield additional candidates (excluding clinical contexts). Conducting a meta-analysis is a laborious process, particularly according to common practice where only a few people do the work, with little support tools and educational materials available. Incentives for creating meta-analyses are low, as public recognition is tied to a single publication. The benefits of meta-analyses for the field, for instance the possibility to conduct power analyses, are often neither evident nor accessible to individual researchers, as the data are not shared and traditional meta-analyses remain static after publication, aging quickly as new results emerge.

A different application of meta-analytic tools is within a paper reporting on several studies. [expand]

3. Report everything and ideally add anonymized supplementary materials. A possible reason for prospective power calculations and meta-analyses being rare lies in the availability of data in published reports. To be able to draw the conclusions we did in this paper, be it in the form of reported effect sizes within paper or as ready-to-use dataset. As noted elsewhere, researchers would ideally always report effect sizes. Despite long-standing recommendations to move beyond the persistent focus on p-values (such as American Psychological Association (2001)), a shift towards effect sizes or even the reporting of them is not (yet) widely adopted (Mills-Smith et al., 2015). A final impediment to meta-analyses in developmental science are current reporting standards, which make it difficult and at times even impossible to compute effect sizes from the published literature.

Best practices when creating new meta-analyses

Data standardization and sharing of materials.

Visualization.

Community-augmented meta-analyses.

Metalab as a model for other domains in child development research.

Conclusion

References

American Psychological Association. (2001). *Publication manual of the american psychological association* (5th ed.). Washington, DC: American Psychological Association.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088–1101.

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *19*(6), 901–917.

Champely, S. (2015). *pwr: Basic Functions for Power Analysis*. Retrieved from https://CRAN.R-project.org/package=pwr

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.

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., ... others. (2004). Clinical trial registration: A statement from the international committee of medical journal editors. Mass Medical Soc.

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1), 1–13.

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

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

Jennions, M. D., & Møller, A. P. (2002). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, *269*(1486), 43–48.

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahnik, S., Bernstein, M., ... others.

(2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76.

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P., Cristia, A., & Frank, M. C. (2016). A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis. *Preprint*. Retrieved from https://osf.io/htsjm/

Mills-Smith, L., Spangler, D. P., Panneton, R., & Fritz, M. S. (2015). A missed opportunity for clarity: Problems in the reporting of effect size estimates in infant developmental science. *Infancy*, *20*(4), 416–432.

Nuijten, M. B., Hartgerink, C. H., Assen, M. A. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547.

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