Building broad-shouldered giants: meta-analytic methods for reproducible research

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

[1] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS

[2] NeuroPsychologie Interventionnelle, ENS

[3] Department Psychology, Stanford University

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

Author Note

Building broad-shouldered giants: meta-analytic methods for reproducible research

Introduction

Psychology has seen a recent "crisis of confidence" in key findings, as many subfields are plagued by issues of low reliability and validity of their data (Ioannidis, 2005). Replicability is a core concept in this recent crisis, as exactly this property of scientific studies and whole fields has come under fire. Replicating a study means 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) with the same outcome as 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 theories can be built on it. In addition, testing different populations and using different, yet comparable stimuli, implies generalizability across supposedly irrelevant dimensions.

Science is a cumulative effort, meaning that a single published report is not sufficient to establish whether an effect is truly present in the population, because misleading reports speaking for or against the existence of a phenomenon might be caused by a number of issues. Next to spurious findings (which can occur even when following best practices due to sampling error alone), a number of habits in psychological research might result in outcomes not reflecting whether or not a phenomenon is truly present in the population. These habits include confining null-results to the filedrawer and running studies with too few participants to reliably detect the effect.

The above mentioned issues are potentially exacerbated in child studies, because the population under investigation is difficult and costly to both recruit and test. Small sample sizes, which in turn lead to habitually underpowered studies. Power, however, is crucial when postulating whether an effect is present and in addition in attempts to estimate its magnitude. In the next section we discuss in detail the possible causes and unwanted negative consequences of underpowered studies, while examining the status quo in developmental research.

A second issue pertains to the way infant researchers gather their data, as all measures are by necessity indirect. In some sub-domains, multiple ways to tap into the same phenomenon have been developed, but are these comparable? We estimate two core measures of interest to infant researchers: drop-out rate and

The present paper aims to quantify both some of the potentially problematic habits of developmental researchers, and show a way forward. We are thus adding to a recently emerging literature that critically examines long-held standards and practices in order to make the whole field more reliable and robust (Mills-Smith, Spangler, Panneton, & Fritz, 2015, Csibra, Hernik, Mascaro, Tatone, & Lengyel (2016)).

To be able to comment on general trends in the field, we make use of a collection of meta-analyses on child development, with a focus on language acquisition. The unique opportunity afforded by such a dataset, which is harnessing data from thousands of participants, is that we can quantify patterns important for experimental practices as well as the role of specific research questions. More precisely, we can quantify whether across different topics, current practices differ. Based on such an assessment, recommendations become possible, that either take the specific of sub-fields into account or can, if supported by the data, be applied across the board.

The Data: MetaLab

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 data.

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 and age range, percent female participants), 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).

Question: More on sampling datasets?

MetaLab contains datasets that address phenomena ranging from infant-directed speech preference to mutual exclusivity, sampled opprtunistically based on data collected with involvement of (some) authors of this paper (n=9 datasets) or they were extracted from published meta-analyses related to language development (n=2, i.e. (Colonessi, **???**)). 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 metalab.stanford.edu and in the supporting information. Further, a throughout investigation into data quality within MetaLab, including publication biases, and a meta-meta-analyses have been conducted based on the same data (Lewis, 2016). It turns out that all databases have evidential value and can thus be utilized for further investigation.

Why power matters. ***General question: Leave here to move to introduction?***

Underpowered studies, that is studies with a low probability to detect an effect given it is present in the population, pose a problem for branches of developmental studies that interpret both significant and nonsignificant findings; for example when tracking the emergence of an ability as children mature or when examining the boundary conditions of an ability. This practice is problematic for two reasons: On one hand, the null hypothesis, for example that two groups do not differ, is not being tested, so it cannot be adopted based on 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. On the other hand, 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.

A Primer on Meta-Analyses. 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, it is possible to compute an estimate of the population effect. Consequently, meta-analyses do not rely on one (possibly false) 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., (Bergmann & Cristia, 2015) for details), for others we estimated this factor based on the information in our database.

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.

Average sample size, effect size, and power per phenomenon

The table below provides summary information for each meta-analysis in MetaLab regarding a number of factors, including the number of single effect sizes and that of papers contributing to a given dataset, in the order of typical age in a specific dataset. The typical sample size as well as the minimum and maximum (allowing to estimate the range in our data) is noted as well. Based on the meta-analytical effect size and the average number of participants, we calculated typical power (using the pwr package (Champely, 2015)). Note 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 [CITE].

*** Question: Is the next bit over-interpreting our data? ***

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. This low number of participants, however, is at least 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.

Table 1
Descriptions of meta-analyses currently in MetaLab.

Topic	Mean Age (Months)	Median Sample Size	Min. Sample Size
Infant directed speech preference	4.50	18.0	10
Sound symbolism	6.77	20.0	11
Vowel discrimination (native)	7.23	12.0	6
Vowel discrimination (non-native)	7.69	16.5	8
Statistical sound category learning	7.98	14.0	5
Word segmentation	8.26	20.0	4
Phonotactic learning	10.69	18.0	8
Label advantage in concept learning	11.49	13.0	9
Gaze following	14.49	24.0	12

Topic	Mean Age (Months)	Median Sample Size	Min. Sample Size
Pointing and vocabulary (longitudinal)	17.98	26.0	12
Online word recognition	18.00	25.0	16
Pointing and vocabulary (concurrent)	22.22	24.5	6
Mutual exclusivity	24.23	16.0	8

REMOVED: Data availability. Awaiting coding for those MAs where it is tracable, will be incorporated in text

TODO: Visualize power

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 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.

Table 2
For each meta-analysis, largest d from oldest paper and power, along with the diffe between power based on meta-analytic and oldest d.

Meta-analysis (MA)	Oldest Paper	Oldest d	Median Sam
Statistical sound category learning	Maye, Werker, & Gerken (2002)	0.56	

Meta-analysis (MA)	Oldest Paper	Oldest d	Median Sam
Word segmentation	Jusczyk & Aslin (1995)	0.56	
Mutual exclusivity	Merriman et al. (1989)	0.70	
Pointing and vocabulary (longitudinal)	Bates et al. (1979)	0.56	
Label advantage in concept learning	Balaban & Waxman (1997)	0.86	
Pointing and vocabulary (concurrent)	Murphy (1978)	0.65	
Phonotactic learning	Chambers et al. (2003)	0.98	
Vowel discrimination (non-native)	Trehub (1976)	1.02	
Sound symbolism	Maurer, Pathman, & Mondloch (2006)	0.95	
Online word recognition	Zangl et al. (2005)	0.89	
Vowel discrimination (native)	Trehub (1973)	1.87	
Infant directed speech preference	Glenn & Cunningham (1983)	2.56	
Gaze following	Mundy & Gomes (1998)	4.52	

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 oldest effect. This difference is larger as oldest effect size increases, with an average of 0.51 compared with an average effect size of 0.59 (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 46.23 participants, instead of 14.07 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.

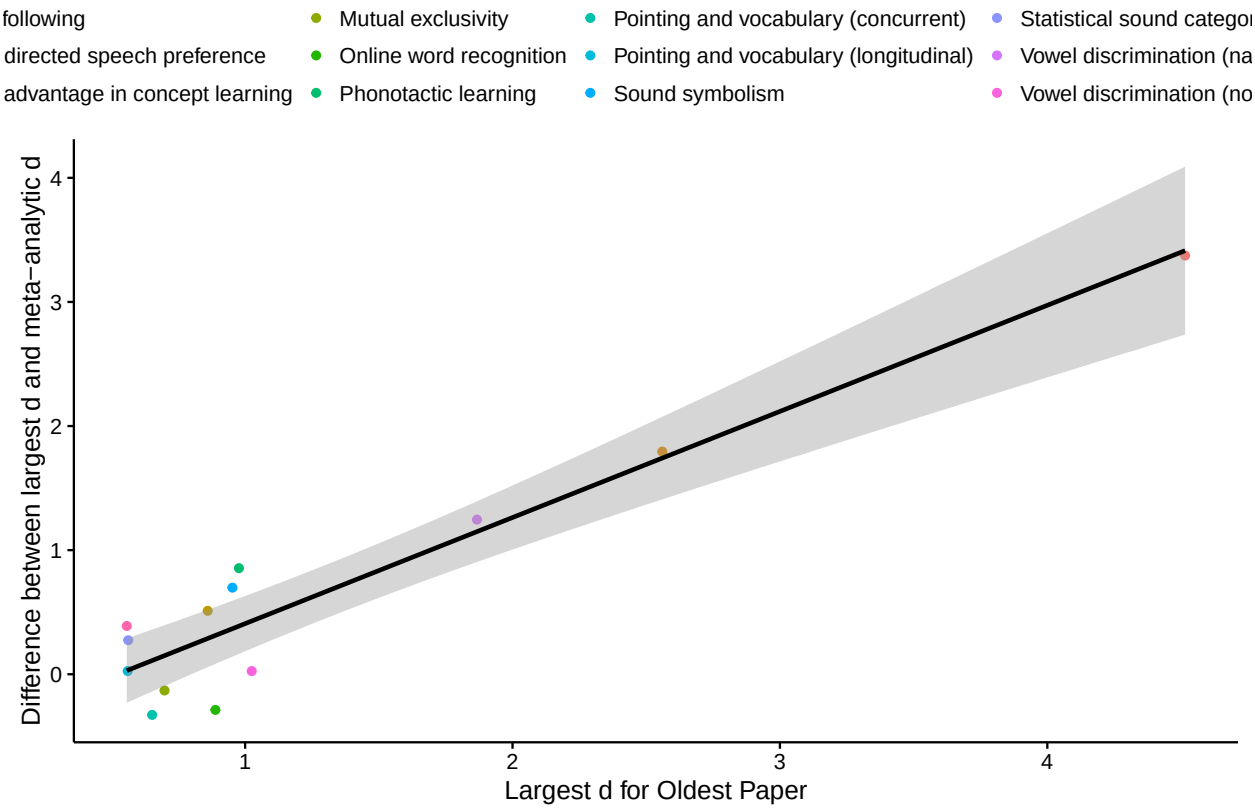

Figure 1. Correlation of largest d from oldest paper and difference between oldest d and meta-analytic d.

Power over time. The comparison of initial and meta-analytic effect size has a number of caveats, for example, as we will lay out in the next section, methods might be different between initial reports and our overall sample; the availability of methods changes over time, as new approaches are being developed and automated procedures become more common. Further, the largest effect size from a seminal paper might have been spurious, and the research community could well be aware of that. In additional, as infant research becomes more common, recruitment and obtaining funds might both become easier, thereby increasing typical sample size over the years. For a more continuous approach, we thus investigate power (which is determined by effect size and sample size) as follows. We first generate a meta-analytic model for each dataset that takes into account infant age and method and then derive the respective to be expected effect size base on those data for each entry in this dataset. Power is then estimated based on the sample size actually tested.

TO DO: Add

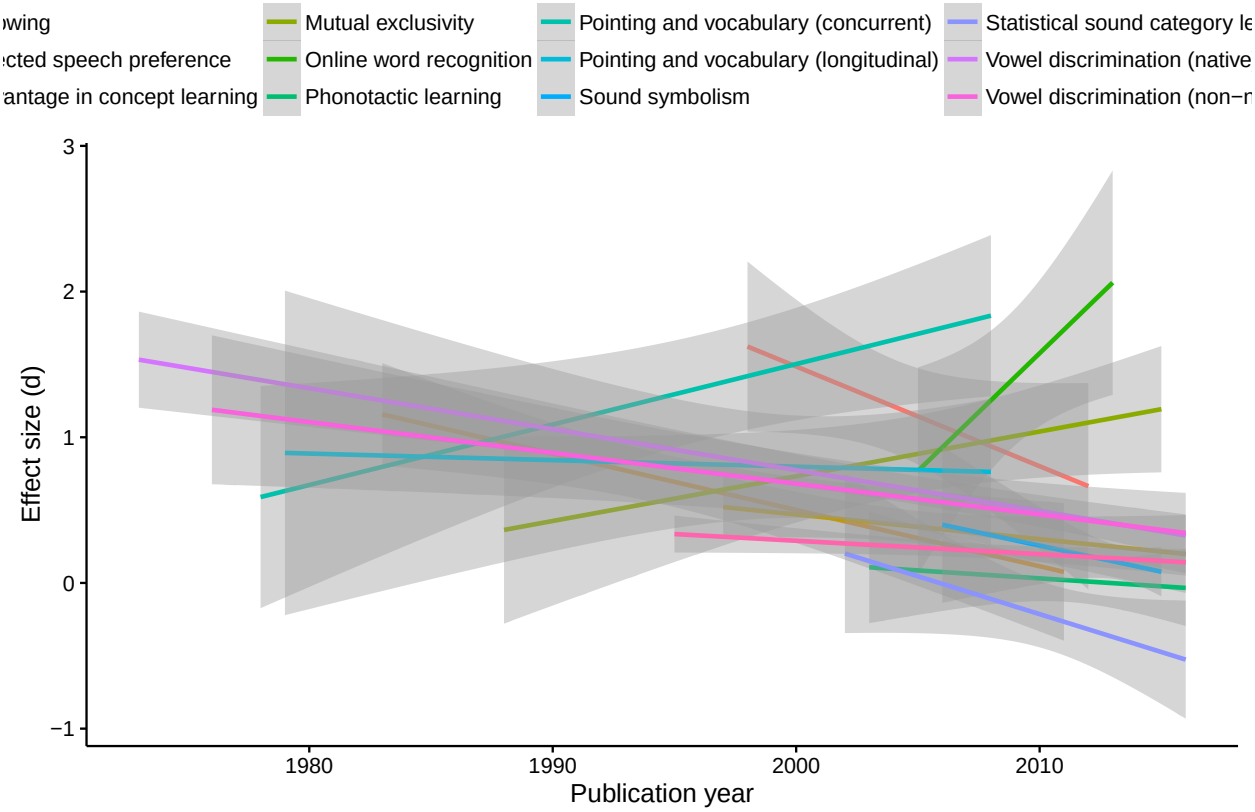

Figure 2. To BE UPDATED: Summary plot of power across years and meta-analyses

How robust are our methods?

Experiment planning goes beyond deciding how many participants to test. Often there is more than one way to measure a specific construct available, even though the number of paradigms available in developmental research when testing infants and children, is somewhat limited by a number of factors, such as time. Consider for example a measurement of preference, such as when trying to establish that infants distinguish infant-directed from adult-directed speech and in fact prefer the former. This preference can be measured in a number of ways, as it is something children bring to the lab. In the meta-analysis on IDS preference there are 4 different methods, all aiming to pick up the very same phenomenon, and this specific line of investigation is no exception, as four datasets of the 11 included datasets contain three or more methods. In this section we address how methods might be

chosen from 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 affected by 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 4 methods and 172 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.

*** Question: method by total participants run (aka ressource-intensity)?***

Table 3
Method vs Dropout

	Estimate	Std. Error	t value
(Intercept)	28.52	2.05	13.93
methodconditioned head-turn	30.44	5.50	5.54
methodhead-turn preference procedure	-2.41	2.85	-0.85
methodstimulus alternation	20.91	3.91	5.35
ageC	0.26	0.45	0.58
methodconditioned head-turn:ageC	2.99	1.15	2.61
methodhead-turn preference procedure:ageC	1.23	0.68	1.81
methodstimulus alternation:ageC	-0.18	0.90	-0.20

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 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.

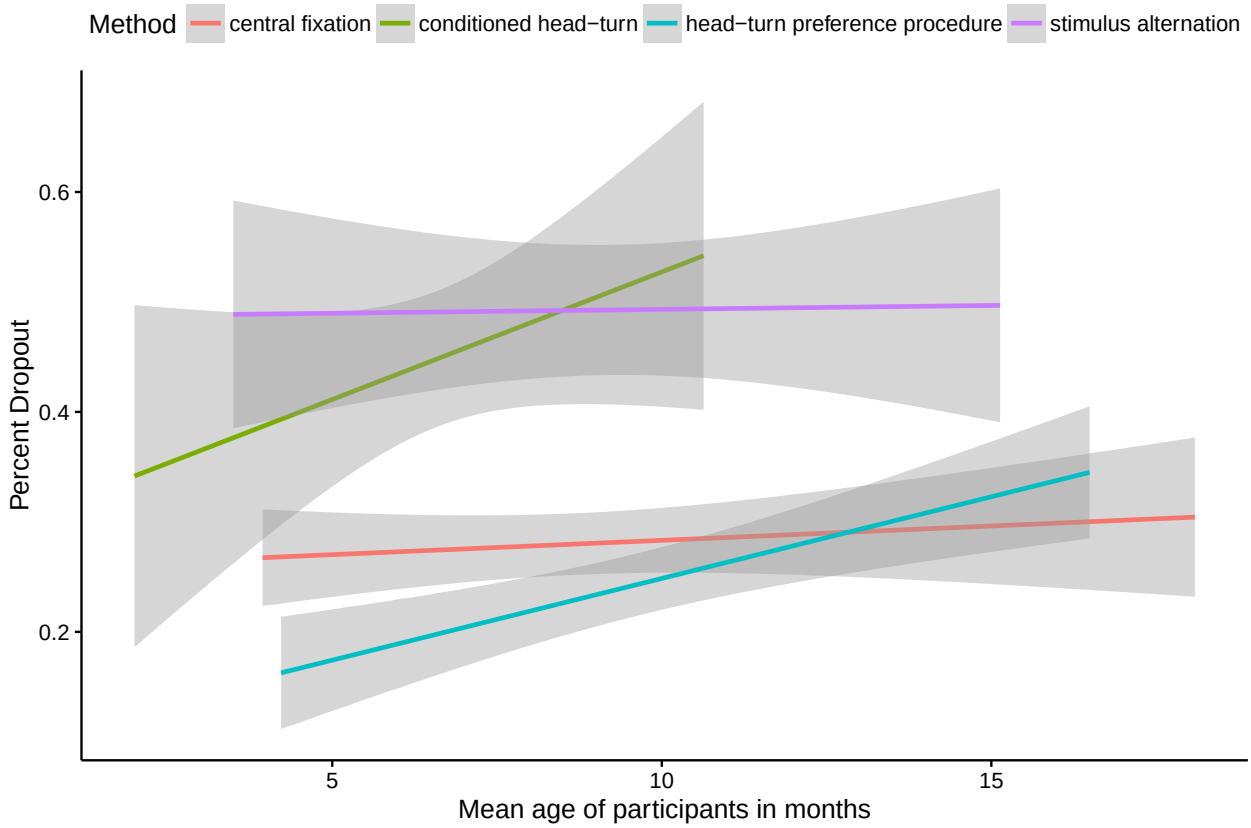

Figure 3. Percent dropout as explained by different methods and mean age of participants.

Table 4
Effect of d by method with central fixation as baseline method.

	estimate	se	zval	pval
intrcpt	0.481	0.107	4.476	0.000
ageC	0.014	0.007	2.073	0.038
relevel(method, “central fixation”)conditioned head-turn	1.137	0.363	3.134	0.002
relevel(method, “central fixation”)head-turn preference procedure	0.000	0.223	0.001	0.999
relevel(method, “central fixation”)high-amplitude sucking	-1.393	3.884	-0.359	0.720
relevel(method, “central fixation”)looking while listening	-0.267	0.213	-1.251	0.211
relevel(method, “central fixation”)stimulus alternation	-0.112	0.207	-0.542	0.588
ageC:relevel(method, “central fixation”)conditioned head-turn	0.075	0.063	1.187	0.235
ageC:relevel(method, “central fixation”)head-turn preference procedure	-0.028	0.027	-1.042	0.297

	estimate	se	zval	pval
ageC:relevel(method, "central fixation")high-amplitude sucking	-0.257	0.449	-0.572	0.567
ageC:relevel(method, "central fixation")looking while listening	-0.009	0.014	-0.625	0.532
ageC:relevel(method, "central fixation")stimulus alternation	0.000	0.028	0.000	1.000

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" 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 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

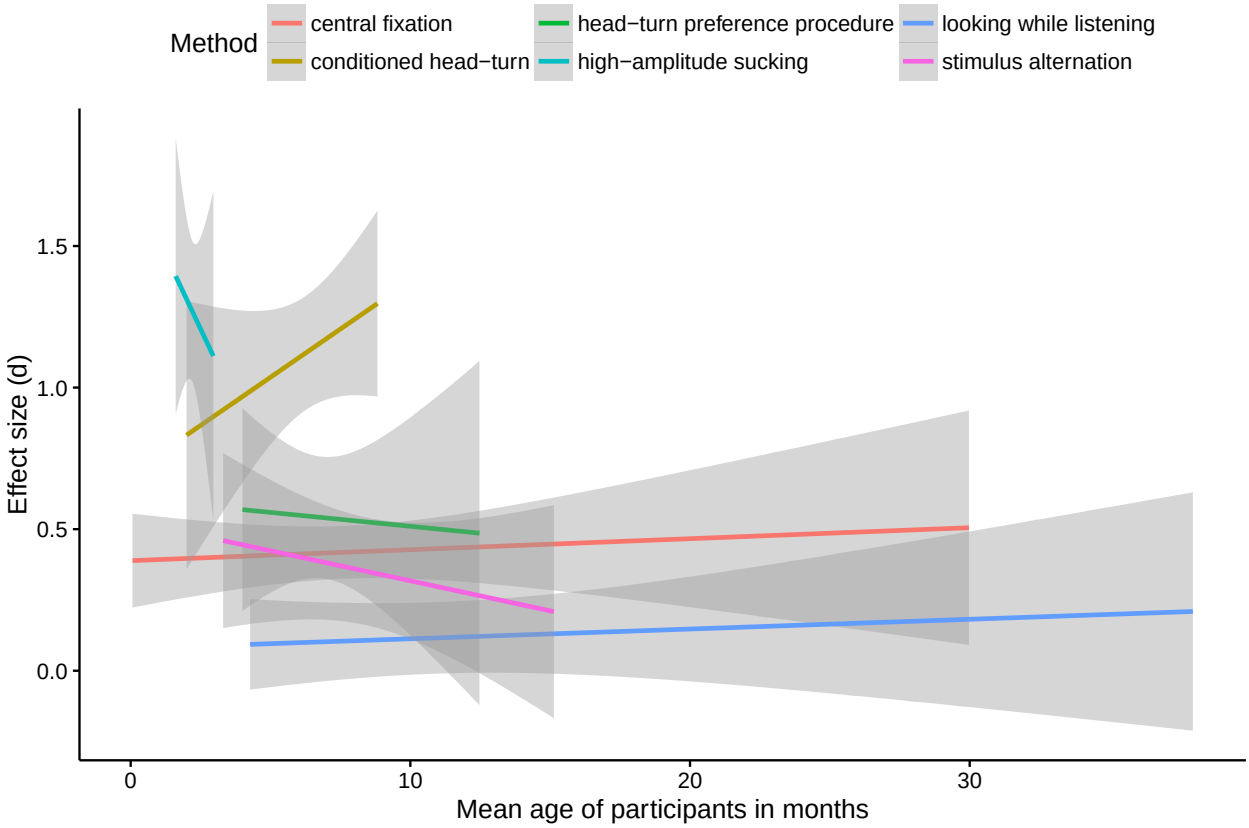

Figure 4. Effect size as explained by different methods and mean age of participants.

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 thus the overestimation is directly reflected in our effect size estimate. We take this caveat into account and compare

Moving forward: Facilitating the creation and use of meta-analyses

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 [cite infancy paper]. 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 12 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 second possible reason lies in the availability of data allowing for the conclusions we were able to draw here, be it in the form of reported effect sizes within paper or as ready-to-use dataset. As noted elsewhere, researchers do not report effect sizes (**???**), despite long-standing recommendations to move beyond the persistent focus on p-values (eg. APA Recommendations). 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.

TODO: ADD data from systematic MAs on author contact

To improve current practices, we propose to address the current issues by making 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 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 (**???**)).

Introducing MetaLab

MetaLab is built on two core principles: lowering hurdles to foster the implementation of practices which are rapidly becoming standard procedure, not only in other branches of psychology (such as effect size estimation and power calculation), and crowdsourcing to decrease the workload of single researchers. MetaLab is based on the recently proposed concept of community-augmented meta-analyses (CAMAs; (Tsuji, Bergmann, & Cristia, 2014)), which combine meta-analyses and open repositories. The advantages of this union are that meta-analyses are shared and get updated continuously, so they can capture the most recent state of the literature and are open to contributions of unpublished results.

MetaLab expands on CAMAs by providing an infrastructure for a range of uses. It is possible to gain an overview of the literature, get insights into specific topics through dynamically rendered reports, conduct power calculations, and contribute not only single recent or unpublished datasets, but whole meta-analyses that can then be opened to contributions and analysis. All meta-analyses share a core of 20 variables which not only allow for the computation of effect sizes across vastly different studies, but also provide the basis for further comparisons. These comparisons are both of practical and theoretical

importance, for example can we compare which method is more robust and suitable for various ages. Researchers then can both better compare existent findings and plan their own research to be more effective. This becomes possible by focusing on a high-level but specific and constrained topic, in the case of MetaLab this is early language development and adjacent phenomena.

MetaLab also poses several advantages compared to existing software for meta-analyses. First, adding meta-analyses is supported not only by sharing standardized formats but also by offering guidance in identifying the correct data to enter, based on the extant data and examples from the developmental psychology literature. Further, novice users can easily engage with the platform to estimate, for example, effect sizes, or decide on sample sizes with a simple interactive tool. Secondly, since all data and scripts are freely and openly available, it is possible to inspect and if needed correct all computations. Errors are thus removed much more swiftly than would be the case for (commercial) software, without losing the benefit of a stable platform. By changing current practices, we aim to increase the reliability of developmental findings and thus the credibility of the field.

Concrete reommendations.

- Reporting: Make it easier to conduct MAs (include correlations, report means and SDs)
- Reporting: Share failures for a more accurate measure, either in MetaLab or on emerging platforms (both allow for anonymzation if you do not want to associate your name for fear of repercussions)
- Study planning: prospective power
- MiniMA if not available in MetaLab: share then

Recommendation: appreciate meta-analyses more

The reluctance to appreciate meta-analyses is evident when comparing citation rates of the initial paper with a meta-analysis on the same phenomenon. Consider the example of infant-directed speech preference, where infants listen longer to speech stimuli showing the typical characteristics of parents talking to their young children. This phenomenon is both theoretically and practically highly relevant and thus receives substantial attention from the field, not the least in a recent large-scale replication attempt (Frank). A meta-analysis on this phenomenon was published in 2012 (**???**), taking 34 studies into account. The oldest paper stems from 1983 [Cite Glenn & Cunningham "What do babies listen to most? A developmental study of auditory preferences in nonhandicapped infants and infants with Down's syndrome."], and the seminal work (measured by the number of citations) was published in 1990 [CITE Cooper Aslin "Preference for infant-directed speech in the first month after birth"]. Comparing these three papers by the number of citations divided by the years since publication (retrieved from google scholar on September 2, 2016) shows that the seminal paper is cited an order of magnitude more every year (on average 24.3 times) than the meta-analysis (2.75 times). This is indicative of practices both when constructing theories and planning experiment: The quantified evidence is under-appreciated, despite providing a number of useful measures, such as effect sizes for different age groups and for various methodological decisions such as stimulus type (synthetic versus natural speech, the own mother versus a stranger, among other things). This is both highly relevant for theories, as the observation of an increased preference for infant-directed speech is a qualitative observation that can allow for more fine-grained hypothesizing. Practically, the information about effect size changes and the impact of method allow for more robust experiment planning and power calculations. We will come back to the issue of power and the impact of considering a seminal paper versus a meta-analysis below. Similar observations hold for other meta-analyses currently available [CITE inphondb, inworddb, others outside language development?].

While anecdotal, this survey showcases current practices and points to one reason for underpowered studies. If authors only consider a single seminal paper to estimate the number of participants necessary, they might habitually run under-powered studies. We show this in dedicated analyses on meta-analytic versus seminal effect size and the resulting typical power in a literature.

With these tools, what do we have to do?

[Tutorial section]

On the individual level:

- How to determine participants: Power calculator, typical N in the field
- How to run the best possible study: Make design choices to have a more robust measure (smaller sample and more power)
- How do I report my data? Best reporting practices (include correlations for within, always report means and SD); and possibly best visualization practices

Further individual benefits:

- Don't despair when a null result occurs, you can still help the community with it
- For replication / training purposes possible to compare ES and select robust ones

On the general level:

- Evidence becomes more reliable
- New evidence can be integrated with previous work directly without much effort
- Complete, unbiased overview of a research literature
  - Identify unexplained variance
  - Where are gaps?
  - Which moderators (do not) affect outcomes

* Examples from published MAs:
    - InWordDB lack of age effect (predicted and strongly assumed in the field)
    - InPhonDB confirmation of diverging effects for native / nonnative, with a quantitative timeline

References

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

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

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536.

Frank, B., M.C. (). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Retrieved from https://osf.io/27b43/

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.

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

Lewis, B., M. (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.

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/

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Psychological Science*, *9*(6), 661–665.

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/