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 recently seen a “crisis of confidence”, as recent findings challenge both the validity of key findings CITE-MANYLABS1 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, CITE-FALSEPOSPSYC,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 CITE-http://www.nejm.org/doi/full/10.1056/nejme048225. The present paper discusses to what extent these issues are present in developmental psychology, provides a quantitative estimation of the prevalence certain problematic practices, and discuss potential solutions.

CB:This last sentence (together with the next section) implies we will look at p-hacking, which we currently don’t and I wonder whether we should.

Relevance of the confidence crisis for developmental psychology

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.

We believe that all of these descriptions are highly relevant to developmental studies, particularly those focusing early and middle childhood. Since the population under investigation is costly to recruit and difficult test, there is a pressure towards small sample sizes. Moreover, populations are intrinsically variable, possibly leading to greater variance and 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.

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 flexibility in data collection and analysis, as arguably this flexibility exacerbates the incidence of false positives CITE-FALSEPOSPSYC. And yet as developmentalists we often "tweak" our paradigms, and explore new ones, in the constant search for more reliable and robust methods. This is because testing infants and children is inherently difficult, as there are both methodological and time constraints. Explicitly instructing participants to respond with a defined behavior, for example, is not possible before the second or third birthday, adn even then, such measures are not always implementable. Further, emerging technologies, such as eye-tracking and tablets, have been eagerly adopted CITE-TABET [?]. As a result of all these factors, multiple ways to tap into the same phenomenon have been developed, but it is an open question to what extent these lead to comparable results. Moreover, although current discussions appear to discourage flexibility, we believe a strong argument can be made for continuous exploration and evaluation of alternative methods, given that another responsibility we have as scientists is to gather data of the highest quality and interpretability. In that sense, sticking to old methods for the sake of comparability would be uneconomical.

The above description would lead one to believe that developmental psychology, particularly that focusing early childhood, should be liable to the issues that have been identified and discussed in other subfields as leading to a crisis in confidence. In the present paper, we seek to provide a more informed overview of a subset of questions that we have identified as key issues.

Key issues

Replication and replicability.

Replicability is a core concept in the recent crisis, as exactly this property of scientific studies (and potentially whole sub-fields) in Psychology 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 peviously 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.

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.

The second bit is only there implicitly

Sample size and statistical power.

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.

In developmental psychology, reasons for sample size decisions are rarely reported, it is thus unnclear 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 ressource limitations determine sample size, as recruitment can be difficult and is very costly.

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. Often there is more than one way to measure a specific construct. Consider for example a measurement of preference, such as when trying to establish that infants distinguish infant-directed from adult-directed speech (IDS and ADS, respectively) and show a strong preference for 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 in developmental psychology.

We will assess in how far the different methods used to test the same construct vary in their sensitivity. Further, taking possible ressource 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. During data collection and analysis, a number of practices can inflate the number of significant p-values, effectively rendering p-values and the notion of statistical significance meaninglss [false pos psy]. First, flexible stopping rules, including adding observations when the 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. *** Check the next statement: imo correct *** In developmental research, this problematic practice further 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 “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.

*** TODO/ Question: test variety in outcome measures,such as longest look vs looking time vs reaction time vs other?***

Publication biases. *** Include? I think it’s sufficiently covered in previous sections***

As mentioned in the general introduction, 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.

*** TODO: Add table with overview and remedies [?] ***

|Problematic practices |Underlying issue |Symptoms |Solution |

|:——————-|:—————–|:————|:—————-| |Small studies |Effect size under |Unpredictable|Prospective power| | |investigation unknown|Outcomes, low|calculations |

Methods

Source data: MetaLab

AC- Not all information provided below seems relevant to the goal of documenting good/bad practices (e.g. % female participants) from the point of view of the reader. Shouldn’t the goal of this section be to provide the minimal relevant info for the reader to understand how we can answer our questions? ChB: tried to shorten, feel free to streamline more, we will have

extensive submat and have the website

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 subfield. 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 subfields 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. (**???**, (**???**))). 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.

**** To be discussed ***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).*** <– AC: I think the question of publication biases should ABSOLUTELY be targeted in this paper, we need to explain what all this means for our field!!!! *** ***ChB: Is that double dipping or so?***

AC- now this just sounds like unsupported bashing, too bad you had to remove the justification. Incidentally, I don't think it's relevant in this paper that is on practices, and not on main effects

ChB: Actually, we want to incentivice making and using MAs here as well, that's for the recommendations part (too be written once we are clear on the results section, as those should be in parallel)

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., (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 across phenomena.

Results and discussion

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 about 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.

*** 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	M
Infant directed speech preference	4.34	20.00	10	
Vowel discrimination (native)	6.54	12.00	6	
Sound symbolism	6.77	20.00	11	
Vowel discrimination (non-native)	7.69	16.00	8	
Statistical sound category learning	8.16	14.75	5	
Word segmentation	8.25	20.00	4	
Phonotactic learning	10.69	18.00	8	
Label advantage in concept learning	11.96	13.00	9	
Gaze following	14.24	23.00	12	
Online word recognition	18.00	25.00	16	
Mutual exclusivity	23.99	16.00	8	

TODO: Visualize power / Question: How?

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 differe between power based on meta-analytic and oldest d.

Meta-analysis (MA)	Oldest Paper	Oldest d	Median Sample
Statistical sound category learning	Maye, Werker, & Gerken (2002)	0.56	
Word segmentation	Jusczyk & Aslin (1995)	0.56	2
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	2
Online word recognition	Zangl et al. (2005)	0.89	2
Gaze following	Mundy & Gomes (1998)	1.29	2
Vowel discrimination (native)	Trehub (1973)	1.87	
Infant directed speech preference	Glenn & Cunningham (1983)	2.39	2

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).

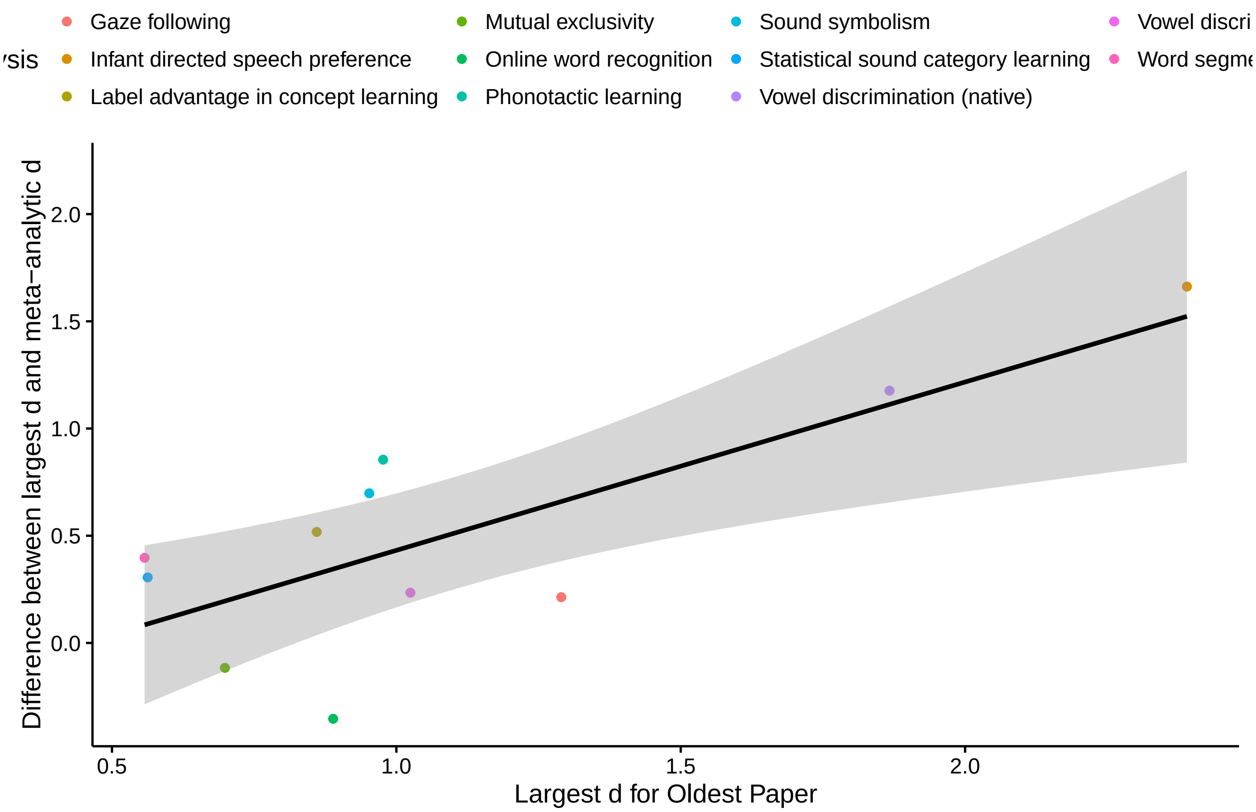

Figure 1. Correlation of largest d from oldest paper and difference between oldest d and meta-analytic d.

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.

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.

Across datasets we observe a general negative trend, with its steepness varyig avross dataset. The only positive trends occur in the upper ranges of estimated power and for older children.

*** Add more here? ***

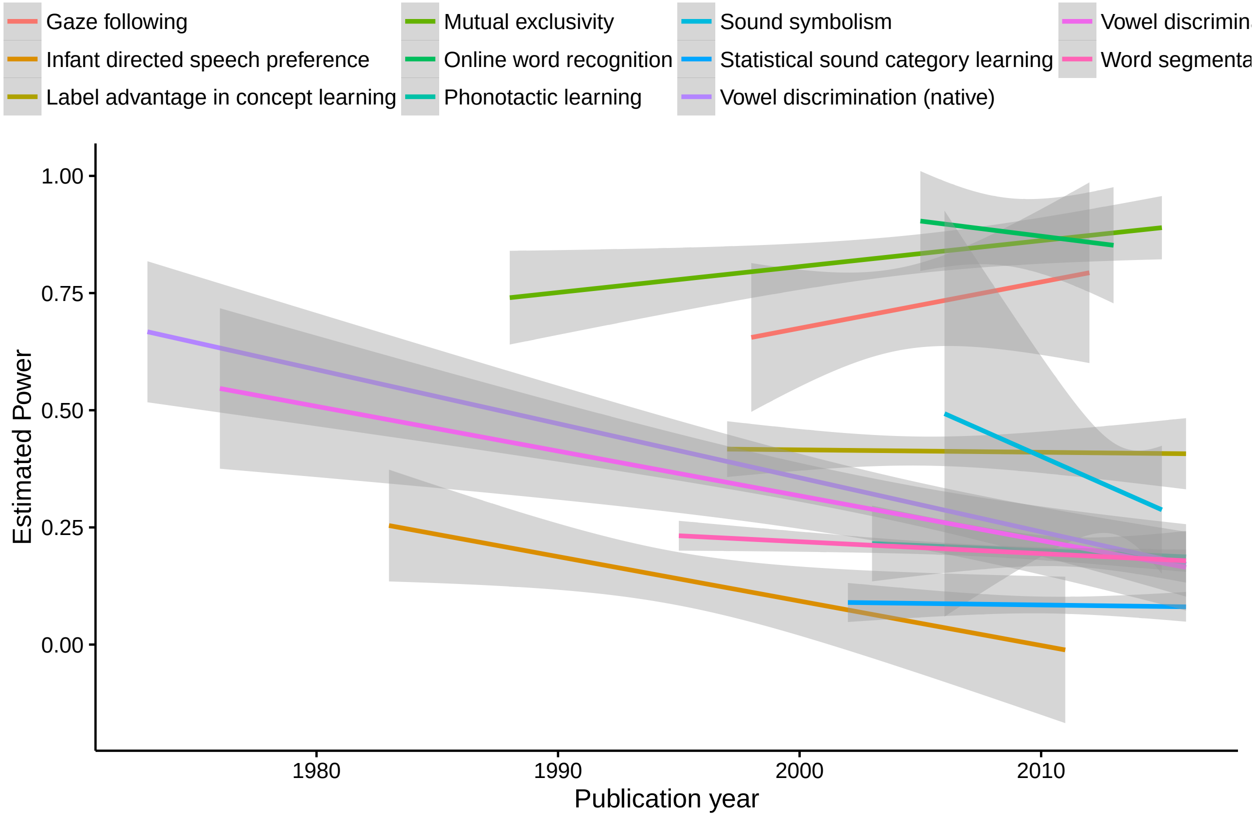

Figure 2. To BE UPDATED: Summary plot of power across years and meta-analyses

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 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 resource-intensity)?***

Table 3
Method vs Dropout

	Estimate	Std. Error	t value
(Intercept)	28.41	2.20	12.90
methodconditioned head-turn	30.20	5.47	5.52
methodhead-turn preference procedure	-2.13	3.22	-0.66
methodstimulus alternation	21.09	3.97	5.31
ageC	0.27	0.47	0.59
methodconditioned head-turn:ageC	2.96	1.18	2.51
methodhead-turn preference procedure:ageC	1.11	0.72	1.54
methodstimulus alternation:ageC	-0.17	0.92	-0.18

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

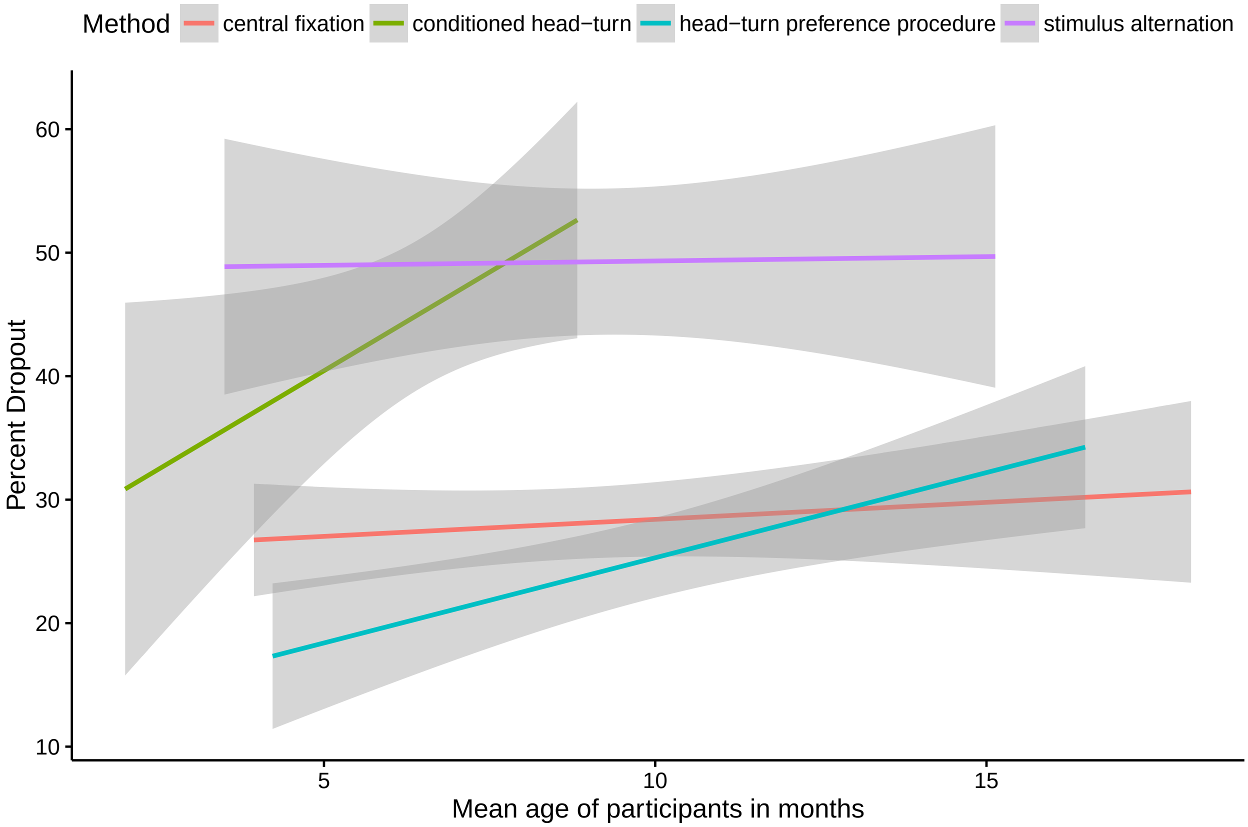

Figure 3. Percent dropout as explained by different methods and mean age of participants.

are limited to 232 observations.

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

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

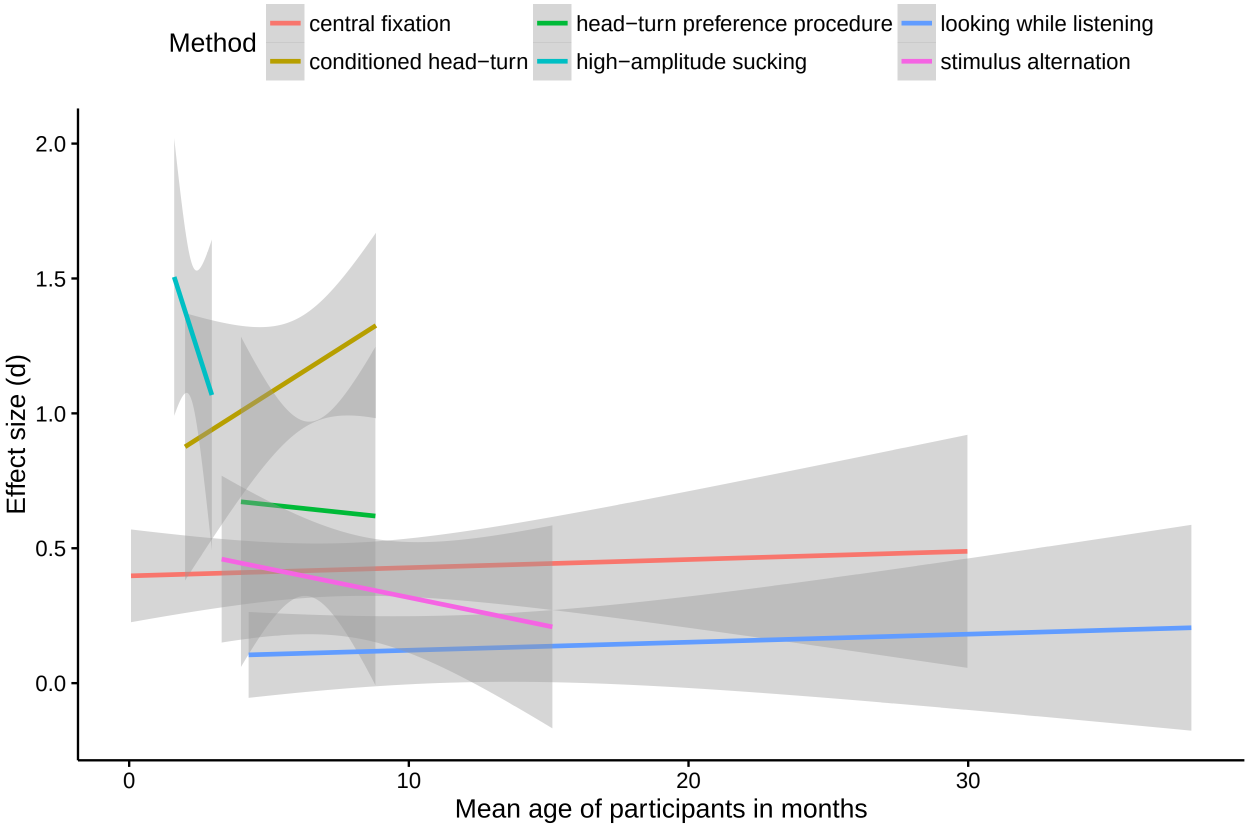

Figure 4. Effect size as explained by different methods and mean age of participants.

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 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

P-hacking and publication biases

*** Note: Previously not part of this paper. Possible analyses include funnel plot asymmetry, p-curves. See synthesis paper. ***

Working towards cumulativity

*** ChB: Alex, what do you envision here?***

Summary and suggestions for the future

We set out to discuss potentially problematic practices. We see X Y and Z. 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)).

Concrete recommendations for developmental psychologists

*** Check table to be made in the beginning ****

1. Calculate power prospectively.

2..

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 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 (**???**)). Finally, by making our data and source code open, we also invite contributions and can update our data, be it by adding new results, filedrawer 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 [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.

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.

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/

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/