# 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 [CITE]. Replicability, that is conducting 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), is a core concept in this recent crisis. 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. This means that a single published report is not sufficient to establish the existence of a phenomenon, and misleading reports might be caused by a number of issues. Next to spurious findings (which can occur even when following best practices), a number of habits in psychological research might result in outcomes not reflecting whether or not a phenomenon is present in the population. These habits include running underpowered studies as well as confining non-significant results to the file-drawer.

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 and noisy measures are a consequence, which in turn lead to habitually underpowered studies. It is thus no surprise that some assume the next "crisis of confidence" brought about by low replicability of core effects will be concerning the field of developmental psychology (Frank).

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. 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.

We explain in the next section in detail the meta-analytical approach and why power is a crucial factor for experimental sciences.

**The Dataset: MetaLab**

The subsequent analyses are based on MetaLab, an online collection of meta-analyses on early 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, 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. Currently, MetaLab contains 12 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, various meta-analyses are combined that address phenomena ranging from infant-directed speech preference to mutual exclusivity. Those datasets were either added by the authors (n=XXX) or extracted from published meta-analyses related to language development (n=2, i.e. (Colonessi, **???**)). In the former case, we attempted to code as much detail as possible for each entered experiment (note that a paper can contain many

experiments). A high level of detail allows not only to retrieve general measures of interest, but also to conduct follow-up analyses into different possible research questions and prospective power calculations taking as much methodological detail as possible into account.

Overall, parts of each meta-analysis are standardized to allow for the computation of common effect size estimates and for analyses that span 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 infomation (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).

As dependent measure, we report Cohen's *d*, a standardized effect size based on comparing sample means. This effect size was calculated when possible from sample 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 the variance of this effect size, 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. 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.

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.

**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. Phenomena differ in the age groups typically tested and the age range covered. This is of high importance, both theoretically, as younger infants might generate more noisy behaviors and are not as advanced in their linguistic abilities, and practically, as older infants might be subjected to more robust methods and could be a more readily available participant pool. 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. 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.

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.

Table 1
*Descriptions of meta-analyses currently in MetaLab.*

| Meta Analysis (MA) | Mean Age in Months | Mean Sample Size | Min. Sample Size |
|---|---|---|---|
| Gaze following | 13.63 | 31.61 | 12 |
| Infant directed speech preference | 4.72 | 22.11 | 10 |
| Label advantage in concept learning | 10.96 | 16.44 | 9 |
| Mutual exclusivity | 27.68 | 18.83 | 8 |
| Online word recognition | 20.30 | 39.40 | 16 |
| Phonotactic learning | 10.18 | 19.45 | 8 |
| Pointing and vocabulary (concurrent) | 21.03 | 26.58 | 6 |
| Pointing and vocabulary (longitudinal) | 18.51 | 32.22 | 12 |
| Sound symbolism | 11.76 | 19.02 | 11 |
| Statistical sound category learning | 7.46 | 16.35 | 5 |
| Vowel discrimination (native) | 7.51 | 16.00 | 6 |
| Vowel discrimination (non-native) | 8.08 | 17.69 | 8 |
| Word segmentation | 9.20 | 22.14 | 4 |

**Power: Comparing meta-analytic effect size and oldest paper.** As Table 1 shows, experimenters are habitually not including a sufficient number of participants to observe a given effect, assuming the meta-analytic estimate for a given topic. It might, however, be possible, that power has been determined based on a seminal paper to be replicated. Initial reports tend to overestimate effect sizes [CITE], possibly explaining the lack of power in some sub-domains. We extracted for each dataset the oldest paper and therein the largest reported effect size and re-calculated power accordingly. 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.

Table 2
*For each meta-analysis, largest d from oldest paper and meta-analytic d.*

| Meta-analysis (MA) | Oldest Paper | Oldest d | Mean Sampl |
|---|---|---|---|
| Gaze following | Mundy & Gomes (1998) | 4.52 | |
| Infant directed speech preference | Glenn & Cunningham (1983) | 2.56 | |
| Label advantage in concept learning | Balaban & Waxman (1997) | 0.86 | |
| Mutual exclusivity | Merriman et al. (1989) | 0.70 | |
| Online word recognition | Zangl et al. (2005) | 0.89 | |
| Phonotactic learning | Chambers et al. (2003) | 0.98 | |
| Pointing and vocabulary (concurrent) | Murphy (1978) | 0.65 | |
| Pointing and vocabulary (longitudinal) | Bates et al. (1979) | 0.56 | |
| Sound symbolism | Maurer, Pathman, & Mondloch (2006) | 0.95 | |
| Statistical sound category learning | Maye, Werker, & Gerken (2002) | 0.56 | |
| Vowel discrimination (native) | Trehub (1973) | 1.87 | |
| Vowel discrimination (non-native) | Trehub (1976) | 1.02 | |
| Word segmentation | Jusczyk & Aslin (1995) | 0.56 | |

To illustrate the disparity between the oldest effect size and the meta-analytic effect,

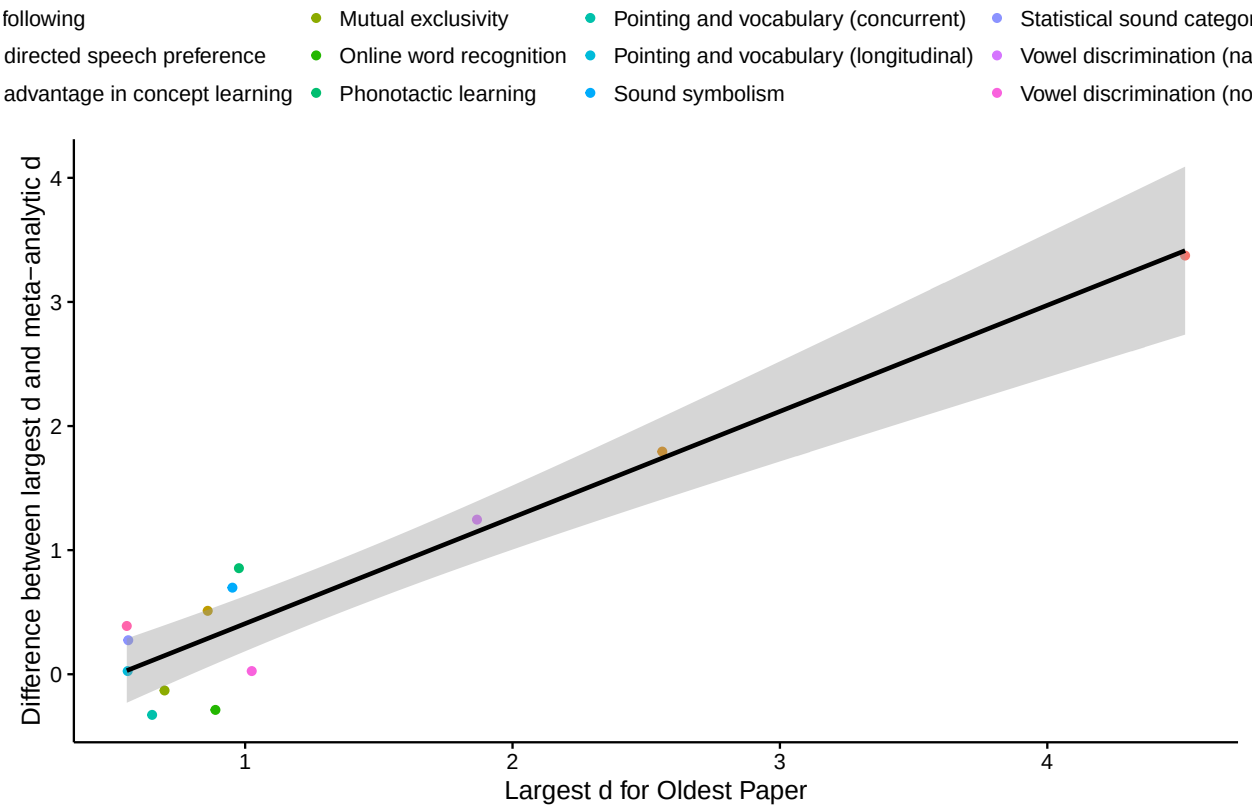


*Figure 1*. Correlation of largest d from oldest paper and difference between oldest d and meta-analytic d.

we plot the difference between both against the oldest effect. This difference is larger as oldest effect size increases, with an average of 0.65 compared with an average effect size of 0.63 (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 40.15 participants, instead of 10.59 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 tp plan for robust and reproducible studies.

**What is the effect of method choice?**

The number of paradigms available in developmental research when testing infants and children, is somewhat limited by a number of factors, such as time. Nonetheless, often there is more than one way to measure a specific construct available. Consider a measurement of preference, for example 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 with this approach this specific line of investigation is no exception, as 4 datasets of the 13 included datasets contain 3 or more methods.

Choosing a robust method can also 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. To this end we consider all methods across datasets in metalab which have more than 10 records. The results of the linear mixed effect model predicting droout rate by method and mean participant age (with dataset as fixed factor to account 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. These effects seem to hold across age groups.

Table 3
*Method vs Dropout*

| | Estimate | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 0.2830431 | 0.0247088 | 11.4551620 |
| methodconditioned head-turn | 0.2693826 | 0.0541791 | 4.9720801 |
| methodhead-turn preference procedure | -0.0249028 | 0.0337799 | -0.7372074 |

| | Estimate | Std. Error | t value |
|---|---|---|---|
| methodstimulus alternation | 0.2161717 | 0.0416199 | 5.1939453 |
| ageC | 0.0026638 | 0.0046998 | 0.5667801 |
| methodconditioned head-turn:ageC | 0.0214482 | 0.0114067 | 1.8803223 |
| methodhead-turn preference procedure:ageC | 0.0122659 | 0.0071521 | 1.7150021 |
| methodstimulus alternation:ageC | -0.0012706 | 0.0093964 | -0.1352257 |

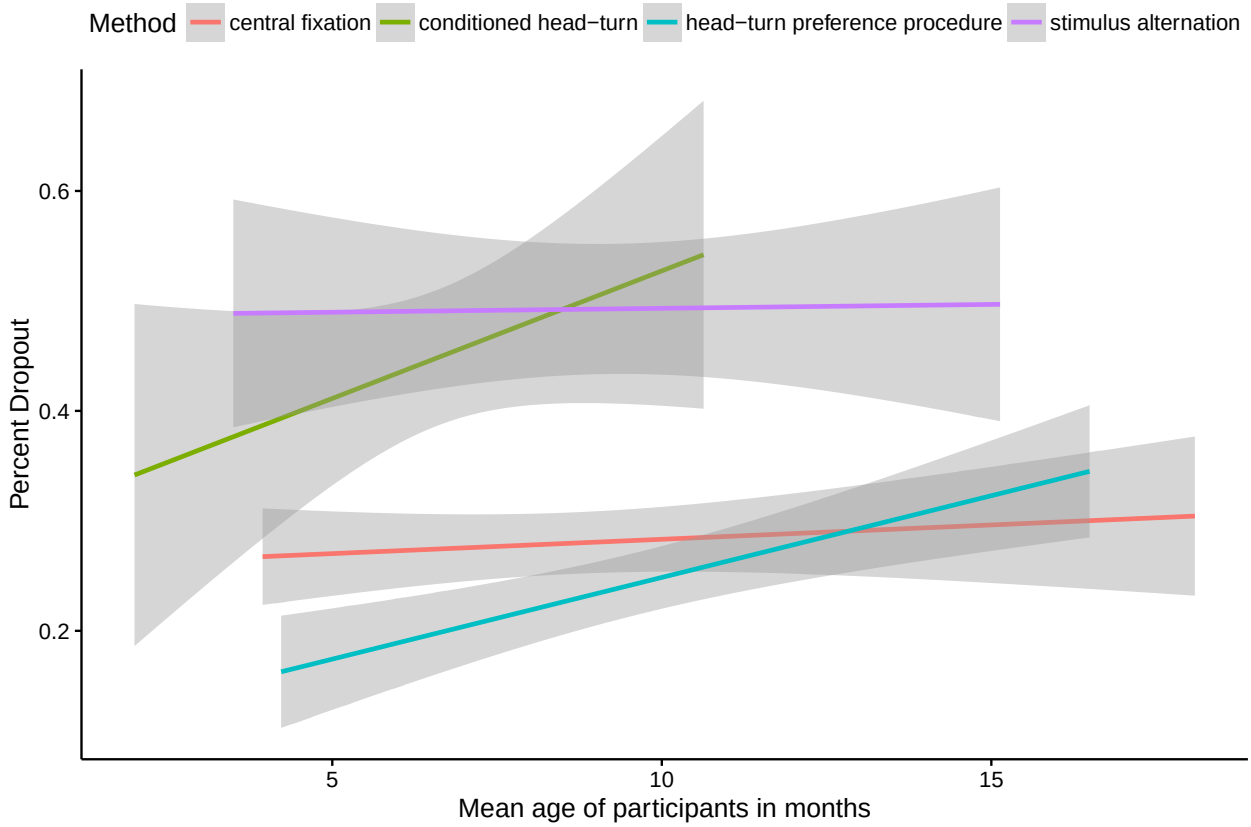


*Figure 2*. Percent dropout as explained by different methods and mean age of participants.

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

| | estimate | se | |
|---|---|---|---|
| intrcpt | 0.5132018 | 0.1156527 | 4. |

| | estimate | se | |
|---|---|---|---|
| ageC | 0.0140214 | 0.0068290 | 2. |
| relevel(method, “central fixation”)anticipatory eye movements | -1.3797138 | 9.8476930 | -0. |
| relevel(method, “central fixation”)conditioned head-turn | 1.6519226 | 0.3462236 | 4. |
| relevel(method, “central fixation”)forced-choice | 1.0215165 | 0.3167291 | 3. |
| relevel(method, “central fixation”)head-turn preference procedure | -0.0250973 | 0.2478096 | -0. |
| relevel(method, “central fixation”)high-amplitude sucking | -1.2312867 | 4.1786021 | -0. |
| relevel(method, “central fixation”)hybrid visual habituation procedure | -4.7403364 | 2.8763087 | -1. |
| relevel(method, “central fixation”)looking while listening | -0.3061340 | 0.2380212 | -1. |
| relevel(method, “central fixation”)oddball | -0.3026225 | 0.4108786 | -0. |
| relevel(method, “central fixation”)stimulus alternation | -0.0496753 | 0.2256594 | -0. |
| relevel(method, “central fixation”)word-object pairing | -0.3321379 | 0.5057796 | -0. |
| ageC:relevel(method, “central fixation”)anticipatory eye movements | -0.4723527 | 3.1999856 | -0. |
| ageC:relevel(method, “central fixation”)conditioned head-turn | 0.1468464 | 0.0592087 | 2. |
| ageC:relevel(method, “central fixation”)forced-choice | -0.0392280 | 0.0211849 | -1. |
| ageC:relevel(method, “central fixation”)head-turn preference procedure | -0.0275691 | 0.0270572 | -1. |
| ageC:relevel(method, “central fixation”)high-amplitude sucking | -0.2315320 | 0.4806216 | -0. |
| ageC:relevel(method, “central fixation”)hybrid visual habituation procedure | -0.8201068 | 0.6222136 | -1. |
| ageC:relevel(method, “central fixation”)looking while listening | -0.0090959 | 0.0145594 | -0. |
| ageC:relevel(method, “central fixation”)oddball | -0.0643442 | 0.0521788 | -1. |
| ageC:relevel(method, “central fixation”)stimulus alternation | 0.0007789 | 0.0288999 | 0. |

We built a meta-analytic model with the effect size measure Cohen's $d$ as the dependent variable, method and mean age of population 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). Only methods with

at least 20 associated effect sizes in MetaLab were included in the model. Thus, the present analyses are limited to 283 observations. 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 (107 times out of 283 total entries of the selected meta-analyses).

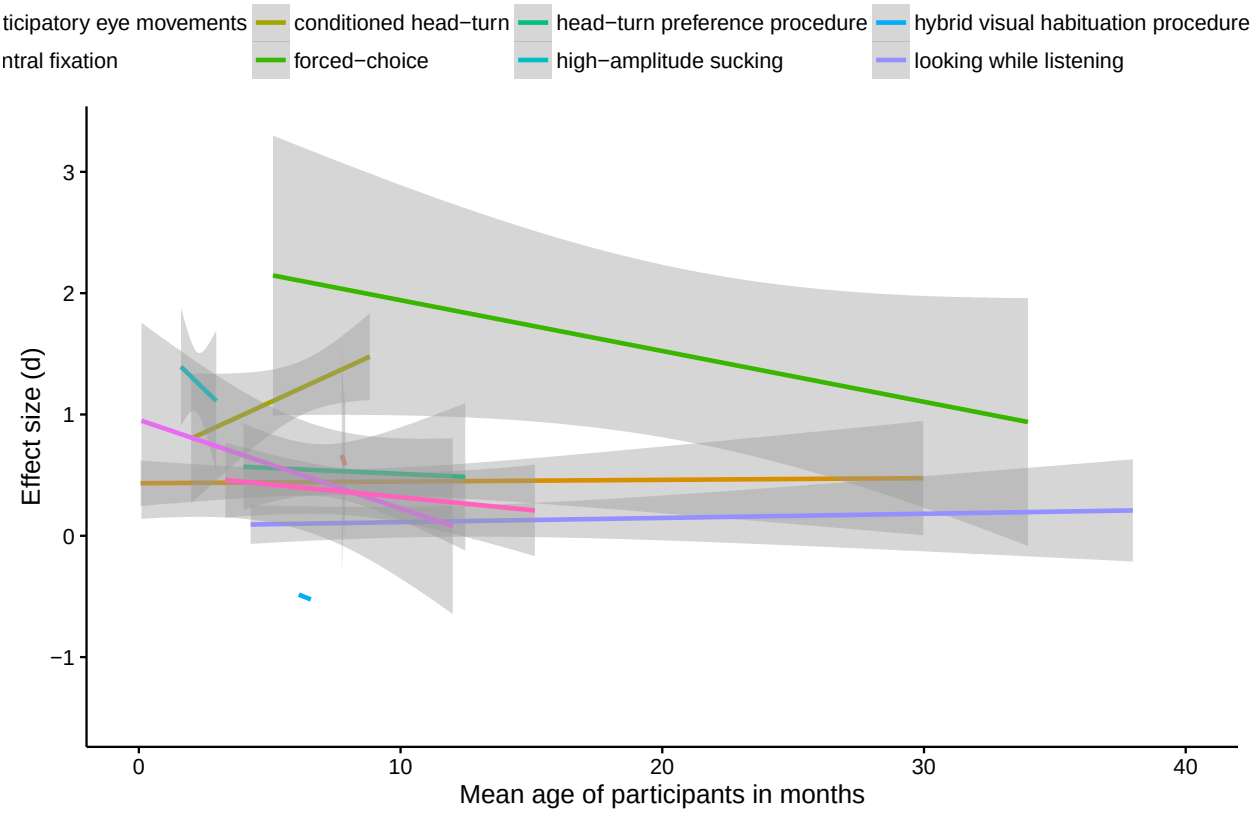


*Figure 3*. Effect size as explained by different methods and mean age of participants.

***TO DO: Add caveats***

## General Discussion

### Recommendation: appreciate meta-analyses more

Meta-analyses and meta-analytic thinking might help solve some of the issues we uncovered.

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 (retreived 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.

**Why don't we do MAs?.** Meta-analyses are also seldomly conducted. This is due to high hurdles and few rewards. Conducting a meta-analysis is a laborious process, particularly according to common practice where only a few people do the work, with little ready-to-use 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 final impediment to meta-analyses in developmental science are, as we illustrated in more detail in section XXX, current reporting standards, which make it difficult and at times even impossible to compute effect sizes from the published literature. As consequence, both systematic, full-scale meta-analyses, and a targeted priori calculation of power and thus the determination of appropriate sample sizes are not yet common practice. Our analyses span various journals and publication years and are thus complementary to recent reports on the overall lack of power in (developmental) psychology based on single-journal/-year samples (e.g., Marszalek, Barber, Kohlhart, & Holmes, 2011).

## Going forward: How can MetaLab help change practices?

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, which has recently come under fire (e.g., Peterson, 2016).

**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*.

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

Frank, M. C. (). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years 1, 2. *Perceptual and Motor Skills*, *112*(2), 331–348.

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.

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius: Sociological Research for a Dynamic World*, *2*, 1–10.

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