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

Author note

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

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)).

The status quo

TO DO: Expand

In this section, we survey current practices in the field of developmental psychology. We focus on reporting and experiment planning practices.

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 papers (n=XXXX). 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. Table XX shows an overview of the phenomena covered.

Average sample size, effect size, and power per phenomenon

A first assessment of current practices in developmental psychology concerns which variables of interest are habitually reported, and what crucial information is ommitted in published papers. To this end we survey the data available in MetaLab, with a focus on participant descriptors, that is age, age range, total participant number and percentage of girls tested; and outcome variables, including the mean and standard deviation of the dependent variable, a test statistic, that is a t-value or F-score, for the main hypothesis test (for example whether looking times differ), and the correlation between outcome variables when infants were tested in multiple conditions (frequently target and control). The data are summarized in Table 1. We break these statistics down by topic and provide the number of papers and experiments (one paper usually contains several experiments). Note that the values are compiled after authors were contacted when information was missing from the published report (details are available in the respective publications). Further, we note whether effect sizes in the form of a variant of Cohen's d were already available; this number includes meta-analyses that were entered based on published papers (**???**, (**???**)).

APA recommendations (American Psychological Association, 2001) have for almost two decades included the requirement to always report and interpret an effect size for all measures of relevance, along with p-values. Effect sizes should, according to those recommendations, be interpreted both when the p-value is above the significance threshold as well as when the significance criterion is met. However, current reproting habits do not follow this recommendation, especially for nonsignificant findings, and if effect sizes are reported, their interpretation is either lacking or misleading (Mills-Smith et al., 2015). This practice has strong implications for theory building, as the data they are based on might not be reliable.

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.

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

Meta Analysis (MA)	Mean Age in Months	Mean Sample Size	Min. Sample Size
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
Statistical word segementation	8.30	21.19	15
Vowel discrimination (native)	7.51	16.00	6
Vowel discrimination (non-native)	8.08	17.69	8
Word segmentation	9.20	22.14	4

TO DO: Exclude dbs based on published MAs from total d reported?!

Table 2
Reporting practices of study outcome measures and demographic information for all papers in MetaLab.

Variable	# Coded	# Uncoded	Total
test_statistic	467	431	898
means	657	241	898
SD	693	205	898
d	228	670	898
corr_within_two	172	358	530
mean_age	898	0	898
age_range	614	284	898

Variable	# Coded	# Uncoded	Total
gender	344	554	898
* TO DO: Add sum	mary acros	s datasets?	***

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 tests into native and non-native vowel discrimination, sample size choices match well with the oldest report.

Table 3
For each meta-analysis, largest d from oldest paper and meta-analytic d.

Meta-analysis (MA)	Oldest Paper	Oldest d	Mean Sampl
Phonotactic learning	Chambers et al. (2003)	0.98	
Vowel discrimination (native)	Trehub (1973)	1.87	
Vowel discrimination (non-native)	Trehub (1976)	1.02	
Pointing and vocabulary (concurrent)	Murphy (1978)	0.65	
Pointing and vocabulary (longitudinal)	Bates et al. (1979)	0.56	
Infant directed speech preference	Glenn & Cunningham (1983)	2.56	
Mutual exclusivity	Merriman et al. (1989)	0.70	
Word segmentation	Jusczyk & Aslin (1995)	0.56	
Statistical word segementation	Saffran, Aslin, & Newport (1996)	-0.39	
Label advantage in concept learning	Balaban & Waxman (1997)	0.86	

Meta-analysis (MA)	Oldest Paper	Oldest d	Mean Sampl
Gaze following	Mundy & Gomes (1998)	4.52	
Statistical sound category learning	Maye, Werker, & Gerken (2002)	0.56	
Online word recognition	Zangl et al. (2005)	0.89	
Sound symbolism	Maurer, Pathman, & Mondloch (2006)	0.95	

*** To Do: align with table 1***

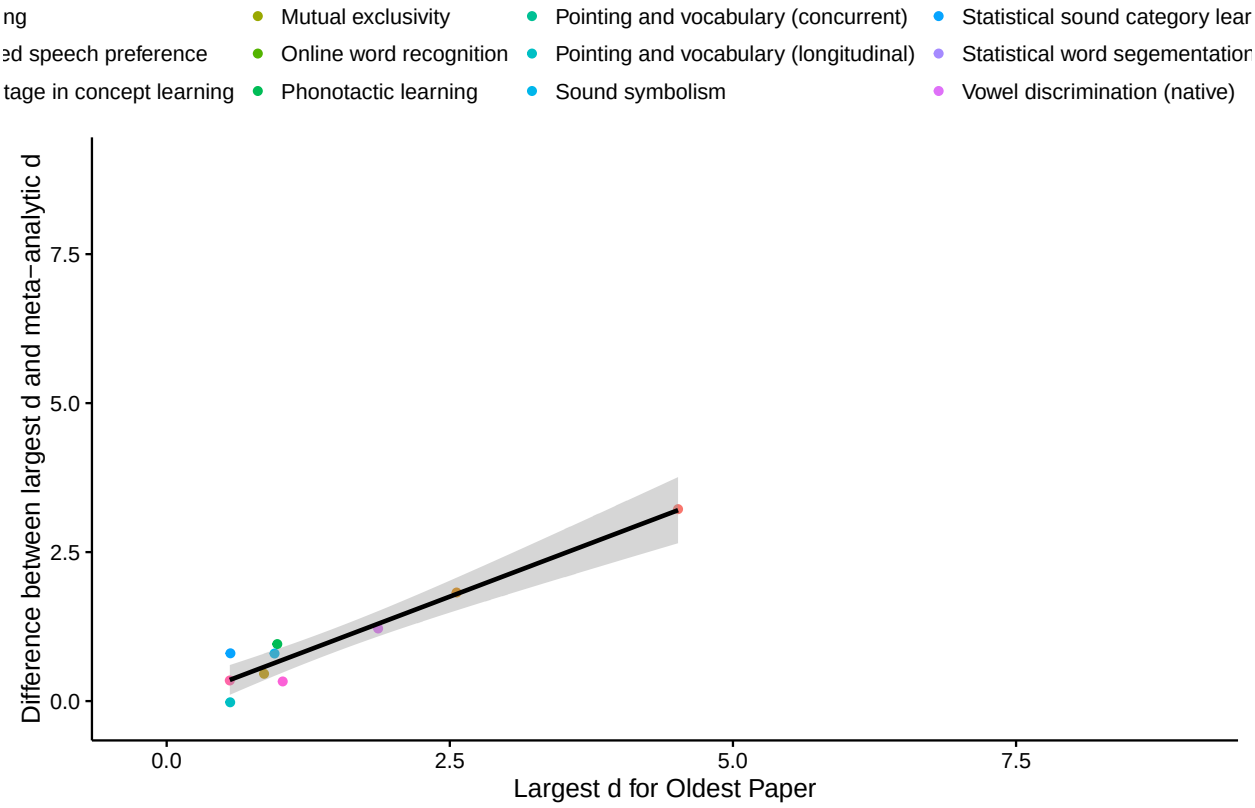

Figure 1. Correlation of largest d from oldest paper and difference between oldest d and meta-analytic d.

To illustrate the disparity between the oldest effect size and the meta-analytic effect, we plot the difference between both against the oldest effect. The difference is larger as oldest effect size increases. This 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.

What is the effect of method choice?

Do researchers chose efficient methods?. Are methods chosen based on the proportion of participants that can be retained for data analysis?

Table 4
Method vs Dropout

	Estimate	Std. Error	t value
(Intercept)	0.3272837	0.0426996	7.664793
methodconditioned head-turn	0.1947841	0.0297706	6.542825
methodhead-turn preference procedure	-0.0437157	0.0357490	-1.222851
methodstimulus alternation	0.2243003	0.0387015	5.795651
ageC	0.0002742	0.0001024	2.678189

Table 5
Effect of d by method with hpp as baseline method.

	estimate	se
intrcpt	0.2892681	0.1421133
ageC	0.0005812	0.0002314
relevel(method, “head-turn preference procedure”)central fixation	-0.1174488	0.0898991
relevel(method, “head-turn preference procedure”)conditioned head-turn	1.6154095	0.3059403
relevel(method, “head-turn preference procedure”)forced-choice	0.1944334	0.1992622
relevel(method, “head-turn preference procedure”)looking while listening	0.1574168	0.2047348
relevel(method, “head-turn preference procedure”)pointing	0.2755598	0.3740201
relevel(method, “head-turn preference procedure”)stimulus alternation	-0.1707422	0.1957474
ageC:relevel(method, “head-turn preference procedure”)central fixation	-0.0001246	0.0003060
ageC:relevel(method, “head-turn preference procedure”)conditioned head-turn	0.0051961	0.0017093
ageC:relevel(method, “head-turn preference procedure”)forced-choice	0.0015039	0.0002833

	estimate	se
ageC:relevel(method, "head-turn preference procedure")looking while listening	0.0009826	0.0004102
ageC:relevel(method, "head-turn preference procedure")pointing	0.0002097	0.0006849
ageC:relevel(method, "head-turn preference procedure")stimulus alternation	-0.0003119	0.0008911

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 865 observations. The included methods are central fixationconditioned head-turnforced-choicehead-turn preference procedurelooking while listeningpointingstimulus alternation. Since the model compares one method as the baseline to all other methods, a baseline method had to be chosen. "Head-turn preference procedure" was included as the baseline method, as it appears most frequently in MetaLab (350 times out of 898 total entries).

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

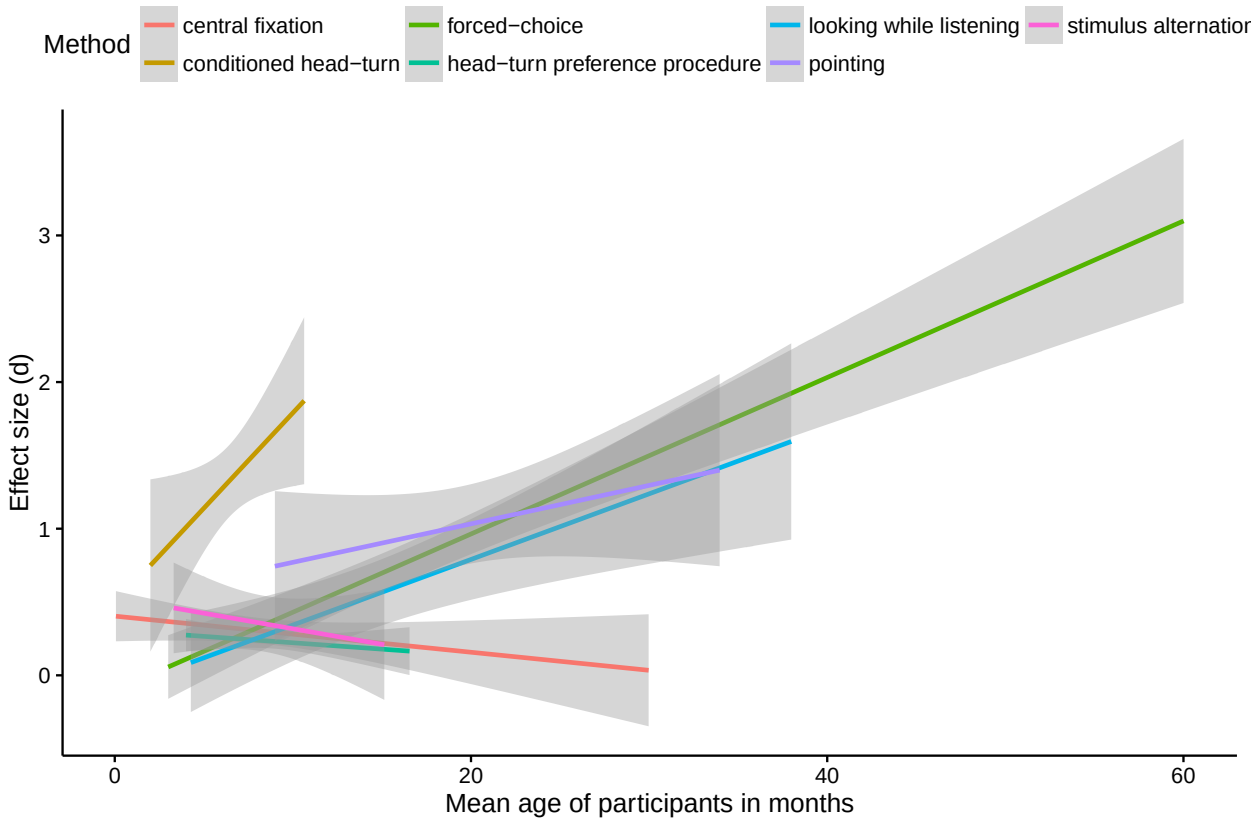

Figure 2. Effect size as explained by different methods and mean age of participants.

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

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

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.