# Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research

Christina Bergmann[1,2], Sho Tsuji[3,1], Page E. Piccinini[4], Molly L. Lewis[5], Mika Braginsky[6], Michael C. Frank[7], & Alejandrina Cristia[1]

[1] LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL Research University
[2] Max Planck Institute for Psycholinguistics, Language Development Department
[3] University of Pennsylvania, Department of Psychology
[4] Neuropsychologie Interventionnelle, Département d'études cognitives, ENS, EHESS, CNRS, PSL Research University
[5] University of Chicago, Computation Institute/University of Wisconsin-Madison, Department of Psychology
[6] Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences
[7] Stanford University, Department of Psychology, Language and Cognition Lab

## Author note

Correspondence concerning this article should be addressed to Christina Bergmann, Max Planck Institute for Psycholinguistics, Language Development Department, PO Box 301, 6500 AH Nijmegen, The Netherlands. E-mail: chbergma@gmail.com

Funding for this research was provided by the Berkeley Initiative for Transparency in the Social Sciences, a program of the Center for Effective Global Action (CEGA), with support from the Laura and John Arnold Foundation. The authors were further supported by the H2020 European Research Council [Marie Sklodowska-Curie grant Nos 660911 and 659553], the Agence Nationale pour la Recherche [ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC], and the Fondation de France.

**Abstract**

Previous work suggests key factors for replicability, a necessary feature for theory building, include statistical power and appropriate research planning. These factors are examined by analyzing a collection of 12 standardized meta-analyses on language development between birth and 5 years. With a median effect size of Cohen's $d$ = 0.45 and typical sample size of 18 participants, most research is underpowered (range: 6%-99%; median 44%); and calculating power based on seminal publications is not a suitable strategy. Method choice can be improved, as shown in analyses on exclusion rates and effect size as a function of method. The article ends with a discussion on how to increase replicability in both language acquisition studies specifically and developmental research more generally.

*Keywords:* replicability, language acquisition, power, study planning, infancy

Word count: 6777

# Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research

Empirical research is built on a never-ending conversation between theory and data, between expectations and observations. Theories lead to new research questions and new data in turn lead to refined theories. This process crucially relies on access to reliable empirical data. Unfortunately, investigators of the scientific process have noted that the assessment of the value of empirical data points can be biased by concerns about publishability (Nosek, Spies, & Motyl, 2012), which often depends on the observation of statistically significant and theoretically surprising outcomes (Sterling, Rosenbaum, & Weinkam, 1995). Aiming for publishability has been suggested to lead to practices that undermine the quality and reliability of data (Ioannidis, 2005; Smaldino & McElreath, 2016). According to some, inappropriate research and reporting practices may be to blame for the surprisingly high proportion of non-replicable findings in psychology (Simmons, Nelson, & Simonsohn, 2011).

Replicability is crucial across domains; but developmental research may be particularly vulnerable to unreliable findings: Collecting data from children is time-consuming, and thus sample sizes are often small, studies are underpowered, and replications are rare. Small sample sizes, and the ensuing lack of power, are a major risk factor for low replicability (e.g., Button et al., 2013). Meta-analysis – the set of statistical tools for aggregating quantitative results across studies – can be a potent tool for addressing issues of replicability. Because no single study is definitive, examining conclusions across studies will facilitate more robust decision-making about the strength of the research literature. In addition, meta-analytic tools can help identify and address issues in replicability by helping to assess weaknesses and allow future studies to be planned more effectively through prospective power analysis. Specifically, a meta-analysis can reveal the average effect size, sample size, and resulting statistical power of a systematically assembled set of studies where a specific phenomenon

has been studied with a variety of methods, stimuli, and samples. Because each meta-analysis typically addresses a single phenomenon – the underlying construct that is supposed to elicit specific responses in laboratory studies – it is difficult to draw general conclusions. To this end, we make use of MetaLab, a publicly available database of 12 standardized meta-analyses of language acquisition. MetaLab is a dynamic, continuously growing database. At the time of writing, the available meta-analyses cover a variety of behavioral and neuroimaging methods (11 in total) and participant ages (from newborns to 5-year-olds).

Since all meta-analyses in MetaLab address specific phenomena within language acquisition, our empirical analyses are adjusted to the methods typically used in this subfield of developmental research. Nonetheless, our analyses and recommendations are relevant beyond the scope of language acquisition research. Crucially, we investigate key study design choices that will be relevant to developmental research at large: sample size (and the ensuing statistical power when effect size is held constant) and method (i.e., paradigms used to tap into the same phenomenon). Furthermore, since our work is comprised of open data and scripts, accompanied by extensive educational materials, and we use open source software (specifically R; R Core Team, 2016), our approach can easily be extended to other domains of child development research. We strongly encourage fellow researchers to build similar collections of meta-analyses describing and quantifying phenomena in their respective sub-domain.

**The meta-analyses in MetaLab**

Before laying out the key concerns for replicability that are more broadly relevant, it may be useful to give a brief overview of our dataset: Each included meta-analysis focuses on one specific phenomenon, and collectively they cover a wide range of linguistic levels, from phonetics (e.g., native vowel discrimination; Tsuji & Cristia, 2014) to pragmatics (e.g.,

pointing and vocabulary; Colonnesi, Stams, Koster, & Noom, 2010) and a range of designs and methods. All but one meta-analysis aggregate experimental studies on the strength of processing of a particular experimentally-manipulated stimulus contrast. The one exception is a meta-analysis containing correlations between toddlers' pointing and vocabulary size measured concurrently (Colonnesi, Stams, Koster, & Noom, 2010). Depending on the meta-analysis and thus phenomenon in question, studies either bear on knowledge acquired outside the lab to tap into continued real-life acquisition processes, or are based on laboratory-based training, typically to isolate a proposed learning mechanism. Examples of the former are native and non-native vowel discrimination (Tsuji & Cristia, 2014) and online recognition of known words (Frank, Lewis, & MacDonald, 2016); the latter is exemplified by learning sound categories and sound sequences in the lab after short exposure to artificial mini-languages (Cristia, 2017). The dependent variable in all these studies is based on continuous response data, such as looking time; either measured within participants in reaction to two conditions or across participant groups receiving different exposures.

Children in our data are aged 0-5 years. In our analyses, we take into account participant age for both practical and theoretical reasons. On the practical side, we expect an effect of infant age based on three aspects of child development research. Firstly, younger infants may be more difficult to recruit and test, thereby increasing measurement noise and leading to smaller effect sizes in younger, compared to older, cohorts. Secondly, tasks and designs might vary as a function of participant age. This factor does not allow us to make a precise prediction with respect to age trends, but does encourage an investigation of research practices and effect sizes as a function of age. Thirdly, even if our tests are conceptually associated to early language acquisition, childhood is a time of rapid cognitive development of various cognitive skills ranging from selective attention to working memory (Lerner,

Liben, & Mueller, 2015), which could impact laboratory performance and would be reflected in the strength, and even direction, of an effect (e.g., Hunter & Ames, 1988).

From a theoretical standpoint, the phenomena targeted by the meta-analyses currently in MetaLab are expected to show changes with age. In general, this change is in a positive direction: Younger participants should show smaller effects than older ones because they are not yet as experienced with, and proficient in, their native language, and thus we expect them to improve in most linguistic skills, such as native vowel discrimination, word form recognition, and word to meaning mapping. The one exception in our collection is non-native vowel discrimination, an ability that should and does *decrease* as infants tune into their native language (Tsuji & Cristia, 2014). For a number of phenomena theoretical predictions are not straightforward (e.g., a preference for infant- over adult-directed speech is thought to increase in the first few months as children accumulate experience with this affective register, but could have been predicted to eventually decrease due to novelty preferences; Hunter & Ames, 1988).

In sum, the set of meta-analysis we use covers a wide range of phenomena and methods, increasing the likelihood that our conclusions are not specific to language acquisition. Moreover, key concerns for replicability, as laid out in the next section, are likely to apply and take effect across sub-disciplines of developmental research. We return to the generalizability of our findings in the discussion.

## Key concerns for replicable research in developmental science

### Statistical power

In this section we review potential hindrances to developmental research being robust and reproducible, and briefly describe how we assess current practices in terms of sampling

decisions and resulting power. All of these descriptions are by necessity brief; for extended discussions we provide references to suitable readings.

In the null-hypothesis significance testing framework, statistical power refers to the probability of detecting an effect and correctly rejecting the null hypothesis if an effect is indeed present in a population. Power is dependent on the underlying effect size and the sample size. Of course, low power is problematic because it increases the likelihood of type-II errors (i.e., failure to find a significant result when there is an effect present in the population). It has become increasingly clear, however, that low power can also increase the frequency of type-I errors (false positives), as the effects reported in such cases will be overestimating the true effect (Button et al., 2013; see also Ioannidis, 2005; Simmons et al., 2011). This fact makes appropriate planning for future research more difficult, as sample sizes will be too small, increasing the likelihood of null results due to insensitive research designs rather than the absence of the underlying effect. In addition, this issue is a serious hindrance to work building on seminal studies, including replications and extensions.

Underpowered studies pose an additional and very serious problem for developmental researchers that interpret significant findings as indicating that a skill is "present" and nonsignificant findings as a sign that it is "absent". In fact, even in the most rigorous study design and execution, null results will occur regularly. Consider a series of studies with 80% power (a number typically deemed sufficient), where every fifth result will be a false negative, that means it will not reflect that there is a true effect present in the population. This observation was recently demonstrated by Oakes (2017) by using data from a high-powered looking time study.

To investigate current practices in our sample, we compute typical power per phenomenon, based on meta-analytic effect sizes and typical sample size (Button et al., 2013).

The logic of this analysis is as follows: Although we cannot know the exact power of any given experiment (because we do not know the true underlying effect), the meta-analytic effect size represents our best guess; thus, the median power for a phenomenon is the power of the median sample size with the meta-analytic effect size. We next explore which effect sizes would be detectable with the sample sizes typically tested in language acquisition research. We additionally investigate how researchers might determine sample sizes using a different heuristic, namely following the largest effect size reported in the first paper on a given phenomenon.

**Method choice**

Improving procedures in developmental research can be considered both an economical and ethical necessity, because developmental populations are difficult to recruit and test. A further complication is that a non-negligible proportion are excluded because they fail to comply, finish the study, or conform to other data quality criteria the researcher sets (e.g., a minimum looking time during test trials). For this reason, developmentalists often "tweak" paradigms and develop new ones with the aim of obtaining a clearer signal and/or control the exclusion rate. Emerging technologies, such as eye-tracking and tablets, have consequently been eagerly adopted (Frank, Sugarman, Horowitz, Lewis, & Yurovsky, 2016; Gredebäck, Johnson, & von Hofsten, 2009).

It remains an open question to what extent the different methods within developmental research lead to comparable results. Some may be more robust, but it is difficult to extract such information based on comparisons of individual studies that use different materials and test various age groups (cf. the large-scale experimental approach by ManyBabies Collaborative, 2017). Aggregating over results via meta-analytic tools allows us to assess to what extent methods differ in their associated exclusion rate as well as to extract general

patterns of higher or lower noise via the comparison of effect sizes since the latter are directly affected by the variance of the measurement.

### Questionable research practices

Undisclosed flexibility during data collection and analysis is a problem independent of the availability of various methods to conduct developmental studies. One salient example is flexible stopping rules, where the decision to stop or continue testing depends on the result of a statistical test. Though this practice might seem innocuous and geared towards "bringing out" an effect the researcher believes is real, it increases the likelihood of obtaining a "significant" outcome well beyond the expected 5%, effectively rendering *p* values and the notion of statistical significance meaningless (Ioannidis, 2005; Simmons et al., 2011).

It is typically not possible to assess whether undisclosed flexibility during data collection (or analysis) led to a false positive in a given report. However, we can measure "symptoms" in a whole literature. We focus in this paper on flexibility in stopping data collection, a practice that was found to be present, but not predominant, in infancy research in a recent anonymous survey (Eason, Hamlin, & Sommerville, 2017). Since our data span over 44 years (publication dates range from 1973 to 2017), it might be the case that recent discussions of best practices have improved lab procedures, but older reports could still have applied this seemingly innocuous practice of adding participants to "bring out" the effect of interest.

## Summary of research goals

We will use a collection of meta-analyses in language acquisition to describe the current state of this field in terms of effect sizes, sample sizes, and, relatedly, statistical power. We take into account the fact that the meta-analyses bear on diverse phenomena, studied in different age groups and with a variety of methods and sample sizes, and that combinations of

these factors will likely affect both effect size and exclusion rates. While we consider the conceptual structure imposed by the fact that the meta-analyses bear on language acquisition, our overarching goal is to exemplify how these analyses can be carried out to describe any subfield of developmental research and to give concrete recommendations and tools to increase replicability within the developmental sciences.

## Methods

All scripts used in this paper, and information how to obtain the source data from MetaLab, are shared on Open Science Framework at https://osf.io/uhv3d/.

### Data

The data presented and analyzed here are part of a standardized collection of meta-analyses (MetaLab), and are freely available via the companion website at http://metalab.stanford.edu. Currently, MetaLab contains 12 meta-analyses, where core parts of each meta-analysis are standardized to allow for the computation of common effect size estimates and for analyses that span across different phenomena. These standardized variables include study descriptors (such as citation and peer review status), participant characteristics (including mean age and native language), methodological information (e.g., 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 of the key hypothesis test, such as *t* values or *F* scores).

Meta-analyses were contributed to MetaLab directly (n=10) or they were extracted from previously published meta-analyses related to language development (n=2; Colonnesi, Stams, Koster, & Noom, 2010; Dunst et al., 2012). In the former case, the meta-analysis authors attempted to document as much detail as possible for each entered experiment (note that a

paper can contain many experiments, as shown in Table 1), as recommended for reproducible and dynamic meta-analyses (Tsuji, Bergmann, & Cristia, 2014). Detailed descriptions of all phenomena covered by MetaLab, including which papers and other sources have been considered, can be found at http://metalab.stanford.edu.

**Statistical approach**

As a dependent measure, we report Cohen's *d*, a standardized effect size based on sample means and their variance. Effect size was calculated when possible from means and standard deviations across designs with the appropriate formulae (Dunlap, Cortina, Vaslow, & Burke, 1996; Lipsey & Wilson, 2001; Morris & DeShon, 2002; Viechtbauer, 2010). When these data were not available, we computed effect size based on the test statistics used to assess the main hypothesis, more precisely *t* values or *F* scores. We also computed effect size variance, which allowed us to weigh each effect sizes when aggregating across studies. The variance is mainly determined by the number of participants; intuitively, effect sizes based on larger samples will be assigned more weight. Note that for research designs testing the same participants in two conditions (for example measuring reactions of 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 (note: publishing guidelines require the reporting of correlations; American Psychological Association, 2001). Some correlations could be obtained through direct contact with the original authors (see e.g., Bergmann & Cristia, 2016). The remaining ones were imputed. We report details of effect size calculation in the supplementary materials and make available all scripts used in the present paper. Excluded as outliers were effect sizes more than three standard deviations away from the median effect size within each meta-analysis (n=12).

### Meta-analytic model

Meta-analytic effect sizes were estimated using random-effect models where effect sizes were weighted by their inverse variance. We further used a multilevel approach, which takes into account not only the effect sizes and 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). When analyzing data from multiple meta-analyses, we nested paper within meta-analysis to account for the fact that studies within meta-analyses will be more similar to each other. We relied on the implementation in the R (R Core Team, 2016) package metafor (Viechtbauer, 2010).

### Power calculation

We calculated typical power using the pwr package (Champely, 2015) based on the meta-analytical effect size and the median number of participants within each meta-analysis. For targeted analyses of the power of seminal papers, we extracted the largest effect size and used this value for power calculation, taking in both cases the median number of participants in a meta-analysis into account (for a similar approach see e.g., Button et al., 2013).

## Results

### Sample size and statistical power

Table 1 provides a summary of typical sample sizes and effect sizes per meta-analysis. We remind the reader that recommendations are for power to be above 80%, which means that four out of five studies show a significant outcome for an effect truly present in the population.

-Insert Table 1 about here-

As could be expected, sample sizes are small across all meta-analyses, with the overall median in our data being 18 infants or paired observations (i.e. 36 participants in total in a between-participant design). Effect sizes predominantly fall into ranges of small to medium effects, as defined by Cohen (Cohen, 1988). The overall median effect size of all data analyzed here is Cohen's $d = 0.45$. As a result of those two factors, studies are typically severely under-powered. Assuming a paired t-test (within-participant designs are the most frequent in the present data), observed power is at 44% (for independent samples, observed power is at 26%).

With the observed sample size, it is possible to detect an effect in 80% of all studies when Cohen's $d = 0.70$; in other words, this sample size would be appropriate when investigating a medium to large effect. When comparing two independent groups, the effect size that would be detectable with a sample size of 18 participants per group increases to Cohen's $d = 0.96$, a large effect that is rarely observed as meta-analytic effect size in the present collection of developmental meta-analyses.

Inversely, to detect the typical effect of Cohen's $d = 0.45$ with 80% power, studies would have to test 40 participants in a paired design; 22 more than are included on average. For a between-participant design, a study with 80% power would require testing 78 infants per group, over four times the typical sample size we encounter here. This disparity between observed and necessary sample size varies greatly across meta-analyses, leading to drastic differences in observed power to detect the main effect. While studies on phonotactic learning and word segmentation are typically dramatically underpowered (with observed power being under 10%), studies on pointing and vocabulary, gaze following, and online word recognition are very well powered (92%, 95%, and 99%, respectively).

We find no strong linear link between participant age and sample size on the level of meta-analyses (Table 1). However, effect sizes and consequently power increase with median participant age. Most saliently, the only three meta-analyses with power over 80%, pointing and vocabulary, gaze following, and online word recognition, typically test participants older than one year.

**Seminal papers as basis for sample size planning**

As Table 1 shows, experimenters only rarely include 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 expanded. Initial reports tend to overestimate effect sizes (Jennions & Møller, 2002), possibly explaining the lack of observed power in the subsequent literature.

For each meta-analysis, we extracted the oldest paper and the largest effect size reported therein and re-calculated power accordingly, using the median sample size of the same meta-analysis (see Table 2). The largest effect size per paper was chosen because many seminal studies contain at least one null result in a control condition that delineates the limitations of a given phenomenon (for example that older children succeed at a task that their younger peers fail). Thus, it is unlikely that the researchers following up on that work aim for the median or mean effect size.

In some cases, such as native and non-native vowel discrimination, as shown in Table 2, 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. In four meta-analyses neither the seminal paper nor meta-analytic effect size seem to be a useful basis for sample size decisions. Since these numbers are based on the largest effect of a

seminal paper, all power estimations (but also differences in meta-analytic effect sizes) would be smaller, meaning that sample sizes are less appropriate than implied by the column denoting power based on the seminal paper in Table 2.

-Insert Table 2 about here-

## Method choice

### Exclusion rates across methods

In most of the analyzed meta-analyses, multiple methods were used to tap into the phenomenon in question. Choosing a robust method can help increase power, because more precise measurements lead to larger effect sizes due to reduced measurement variance and thus require fewer participants to be tested to conduct appropriately-powered studies. However, the number of participants relates to the final sample and not how many participants had to be invited into the lab. We thus first quantify whether methods differ in their typical exclusion rate, as economic considerations might drive method choice. To this end we consider all methods which have more than 10 associated effect sizes and for which information on the number of excluded participants was reported and entered in the meta-analyses. We note that this is exclusion rate, rather than fussout or dropout rates, because it represents the number excluded considering all criteria, including data quality criteria such as a minimum looking time. We chose this variable for practical reasons, as overall exclusion rates are more frequently reported than the number of participants who did not complete the experiment. The following analyses cover 6 (out of 11) methods and 224 (out of 761) effect sizes.

The results of a linear mixed effects model predicting exclusion rate by method and mean participant age (while controlling for the different underlying effect sizes per meta-analysis) are summarized in Table 3 and visualized in Figure 1. The results show significant

variation across methods, and a tendency toward higher exclusion rates for older participants, with some interaction with method.

-Insert Table 3 about here-

-Insert Figure 1 about here-

**Effect sizes as a function of method**

We built a meta-analytic model with Cohen's *d* as the dependent variable, and method and mean age centered as independent variables, which we allowed to interact. The model includes the variance of *d* for sampling variance, and a nested random effect of paper (inner random effect) within meta-analysis (outer random effect). We limited this analysis to the same methods that we investigated in the section on exclusion rates to be able to observe possible links between effect size and exclusion rate in methods. The model results in Table 4 show significant variation in effect sizes across methods, age, and some interaction of method and age.

-Insert Table 4 about here-

-Insert Figure 2 about here-

**Questionable research practices**

In the final set of analyses, we assess the relation between absolute observed effect sizes in single studies and the associated sample size. The rationale behind this analysis is simple: The smaller the effect size in a particular study (bear in mind that we assume that experiments sample from a distribution around the population effect), the larger the sample needed for a significant *p* value. If sample size decisions are made before data collection and all results are published, we expect no relation between observed effect size and sample size. If, on the

contrary, authors continue to add infants to achieve significance (Begg & Mazumdar, 1994), there should be a negative correlation between sample size and effect size.

-Insert Figure 3 about here-

We illustrate the link between effect size and sample size, separated by meta-analysis, in Figure 3. The statistical test results for each meta-analysis can be found in Table 5. Four meta-analyses show a significant negative relation between sample size and effect size, consistent with bias; two of them assess infants' ability to discriminate vowels, one bears on word segmentation, and one tests whether children use mutual exclusivity during word learning. The last case might be driven by a single high-powered study with an atypical developmental range (Frank, Lewis, & MacDonald, 2016). We further observe an unexpected positive correlation between sample size and observed effect size in the meta-analysis on infant directed speech preference, which we discuss below.

-Insert table 5 about here-

## Discussion

In this paper, we made use of a collection of 12 standardized meta-analyses to assess typical effect sizes, sample size, power, and methodological choices that are currently common in research on language development. With a median effect size of Cohen's $d = 0.45$ and a typical sample size of 18 participants per cell, observed power is only 44%.

The lack of power is particularly salient for phenomena typically tested on younger children, because sample sizes and effect sizes are both small (the one exception for research topics tested mainly with participants younger than one year is non-native vowel discrimination, which can be attributed to a large meta-analytic effect size estimate rather than larger samples). Phenomena studied among older children tended to yield larger effects, and

here some studies turn out to be high-powered (e.g., online word recognition). Both observations are first indicators that effect size estimates might not be considered when determining sample size, as power of 99% would suggest the sample was unnecessarily large for the effect under study (see Table 1). However, it is possible that, in addition to testing a main effect (such as whether children recognize a given word online) these high-powered studies also investigated interactions (i.e., factors modulating this ability). As a consequence, studies might be powered appropriately since an interaction effect will be more difficult to detect than a main effect. The possibility that follow-up studies are looking for moderators and thus test interaction effects means that the 44% average power observed above would be an overestimate.

We next investigated the possibility that researchers base their sample size on the highest effect size reported in the seminal paper of their research topic. We find that even under this assumption, the surveyed research would largely be underpowered. Moreover, this strategy would likely not provide sufficient power with respect to meta-analytic effect sizes, as early explorations will tend to overestimate effect sizes (Jennions & Møller, 2002). In short, studies are habitually underpowered because sample sizes typically remain close to what can be called a "field standard" of 15 to 20 participants (see Table 1 in this paper and Oakes, 2017).

Conducting studies with sample sizes based on "field standards" is highly problematic for several reasons. First, many studies will not yield significant outcomes despite the presence of a real, but small effect. Researchers might thus be inclined to conclude that an ability is absent in a population (see below for an in-depth discussion of this topic), or they may refrain from publishing their data altogether. If an underpowered study is published because the outcome is significant, this study will overestimate the size of the underlying effect, thereby adding biased results to the available literature (and thus further biasing any

meta-analytic effect size estimate; Sterling et al., 1995; Yarkoni, 2009), as well as reinforcing the practice of sampling too few participants. At worst, this practice can lead to the perpetuation of a false hypothesis (for an example, albeit from non-developmental research, consider the meta-analysis of romantic priming by Shanks et al., 2015).

We investigated the possibility that researchers selectively add participants to obtain a significant result through the relation between observed effect size and sample size. We observed that in four meta-analyses effect sizes were significantly negatively correlated with sample sizes, which might be an indication of questionable research practices. At the same time we found a (numerically) positive correlation in the meta-analysis on infant-directed speech preference, an unexpected result as it means that larger sample sizes tend to be found in experiments with larger effects. One possible reason for the latter result might be specific to this dataset: perhaps older infants are both easier to test and have greater preferences for infant-directed speech.

For the four observed negative correlations, alternative explanations to questionable research practices are possible: As soon as researchers are aware that they are measuring a more subtle effect and adjust sample sizes accordingly, we expect to observe this negative correlation. Consider for example vowel discrimination, which can be studied with very distinct vowel-pairs such as in "bit" and "but", or with subtler contrasts like in "bat" and "bet". In fact, in the presence of consequent and accurate a priori power calculations, a negative correlation between sample size and effect size must be observed. However, our previous analyses indicate that power is not considered when making sample size decisions.

**Concrete recommendations for developmental scientists**

In this section, we move from a description of current practices to suggestions aimed at improving the reproducibility of developmental research. We generalize to developmental

studies at large because there is reason to believe that other sub-domains in the study of infant and child development may be subject to the same issues we outlined in the introduction.

### 1. Calculate power prospectively

We found that most studies testing infants and toddlers are severely underpowered, even when aiming to detect only a main effect. Interactions will show smaller effect sizes and thus will be even harder to detect. Further, power varies greatly across phenomena, which is mainly due to differences in effect sizes. Sample sizes are not adjusted accordingly, but remain close to the typical sample size of 18.

Our first recommendation is thus to assess in advance how many participants would be needed to detect a minimal effect size of interest (for a more detailed discussion and practical recommendations see Lakens & Evers, 2014). Note that we based our power estimations on whole meta-analyses, an analysis approach most suitable to making general statements about a research field at large. It might, however, be the case that specific studies might want to base their power estimates on a subset of effect sizes to match age group and method. Both factors can, as we showed in our results, influence the to-be-expected effect size. To facilitate such analyses, all meta-analyses are shared on MetaLab along with the available details about procedure and measurements (see also Tsuji et al., 2014).

In lines of research where no meta-analytic effect size estimate is available – either because it is a novel phenomenon being investigated or simply due to the absence of meta-analyses – we recommend considering typical effect sizes for the method used and the age group being tested. This paper is a first step towards establishing such measures, but more efforts and investigations are needed for robust estimates (Cristia, Seidl, Singh, & Houston, 2016).

**2. Carefully consider method choice**

One way to increase power is the use of more sensitive measurements; and we do find striking differences between methods. On the practical side, exclusion rates varied a great deal (with medians between 5.9% and 45%). Interestingly, the methods with somewhat lower exclusion rates (central fixation and headturn preference procedure) are among the most frequent ones in our data. The proportion of participants that can be retained might thus inform researchers' choice. This observation points to the previously mentioned limitations regarding the participant pool, as more participants will have to be tested to arrive at the same final sample size. High exclusion rates can also be offset by high effect sizes; as can be seen when comparing conditioned headturn in Figures 1 and 2, while exclusion rates are around 30-50%, effect sizes are above 1. The second method with high exclusion rates, stimulus alternation, in contrast, does not fall into this pattern of high exclusion rates coinciding with high effect sizes. A possible interpretation of this finding is that some methods, which have higher exclusion rates, generate higher effect sizes due to decreased noise (e.g., by excluding participants who are not on task). However, there is an important caveat: Studies with fewer participants (thanks to higher exclusion rates) are imprecise, and thus it is more likely that significant results overestimate the underlying effect.

Nevertheless, when possible, it seems important to consider the paradigm being used, and possibly use a more sensitive way of measuring infants' capabilities. One reason that researchers do not appear to choose the most robust methods might again be due to a lack of consideration of meta-analytic effect size estimates, which in turn might be (partially) due to a lack of information on (how to interpret) effect size estimates and lack of experience using them for study planning (Mills-Smith, Spangler, Panneton, & Fritz, 2015). We thus recommend to change this practice and take into account the possibility that different methods’ sensitivity is reflected in effect size. Efforts to estimate the impact of method choice

experimentally through large-scale replications will likely be informative in this quest (Frank et al., 2017).

### 3. Report all data

A possible reason for prospective power calculations and meta-analyses being rare lies in the availability of data in published reports. Despite longstanding recommendations to move beyond the persistent focus on *p* values (such as American Psychological Association, 2001), a shift towards effect sizes or even the reporting of them has not (yet) been widely adopted (Mills-Smith et al., 2015).

In addition, in cases where effect sizes are not mentioned, current reporting standards make it difficult – at times even impossible – to derive effect sizes from the published literature. For example, for within-participant measures it is necessary to report the correlation between repeated measures associated to the paired conditions (most commonly a treatment and control condition). However, this correlation is habitually not reported and has to be obtained via direct contact with study authors (see for example Bergmann & Cristia, 2016) or estimated (as described in Black & Bergmann, 2017). In addition, reporting (as well as analysis) of results is generally highly variable, with raw means and standard deviations not being available for all papers.

We suggest reporting the following information, in line with current guidelines: means and standard deviations of dependent measures being statistically analyzed (for within-participant designs with two dependent variables, correlations between the two should be added), test statistic, exact *p* value (when computed), and effect sizes (for example Cohen's *d* as used in the present paper) where possible. Such a standard not only follows extant guidelines, but also creates coherence across papers and reports, thus improving clarity (Mills-Smith et al., 2015). A step further would be the supplementary sharing of all

anonymized results on the participant level, thus allowing for the computations necessary for meta-analyses, and opening the door for other types of cumulative analyses.

### 4. Increase the use and availability of meta-analyses

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. The workload associated with conducting a meta-analysis may thus appear (and perhaps even be) much larger than that associated with a publication containing original data or with a qualitative review, making meta-analyses less attractive than the latter two for individuals. Moreover, the benefits of meta-analyses for the field, for instance the possibility of conducting 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 (Tsuji et al., 2014).

To support the improvement of current practices, we propose making meta-analyses available in the form of ready-to-use online tools, dynamic reports, and as raw data. These different levels allow researchers with varying interests and expertise to make the best use of the extant records on language development, including study planning, by choosing robust methods and appropriate sample sizes. An additional advantage of using meta-analysis when interpreting single results is that researchers can easily check whether their result falls within the expected range of outcomes for their research question – indicating whether or not a potential moderator influenced the result.

Meta-analyses can also be useful for theory building. Indeed, aggregating over many data points allows us to trace the emergence of abilities over time, as well as quantify their growth, and identify possible developmental trajectories. A demonstration is given in the work of Tsuji and Cristia (2014), where mainstream descriptions of development for native

and non-native vowel discrimination could be confirmed. Contrastingly, Bergmann & Cristia (2016) showed that word segmentation from native speech does not follow the typically assumed developmental trajectory (for a recent discussion of both meta-analyses see Bergmann, Tsuji, & Cristia, 2017). As a consequence, meta-analytic investigations lead to more refined, or even reconsidered, theoretical accounts of child development, bolstered with a better estimate of the timeline for phenomena of interest (see also Lewis, et al, 2016).

### 5. Use cumulative evidence to decide whether skills are "absent" or not

Developmental researchers often interpret both significant and nonsignificant findings, particularly to establish a timeline tracing when skills emerge. This approach is problematic for multiple reasons, as we mentioned in the Introduction. Disentangling whether a nonsignificant 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 in fact impossible based on *p* values. Further, we want to caution researchers against interpreting the difference between significant and nonsignificant findings without statistically assessing it first (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). As mentioned, meta-analyses provide a more principled way for assessing statistically whether age explains significant proportions of the variance in observed effects. Moreover, this technique can also help with cases where the absence of an effect is incorrectly inferred from a string of nonsignificant, potentially underpowered, studies, as recently demonstrated by Vadillo, Konstantinidis, and Shanks (2016). In their study, the authors pooled null results that had been taken as evidence for an absent effect, and demonstrated the meta-analytic effect size estimate was Cohen's *d* = .3 (an effect that happens to be larger than that found in some meta-analyses included here).

## Future directions

The present analyses can be expanded and improved in a number of ways. First, this collection of meta-analyses does not represent an exhaustive survey of phenomena in language acquisition, let alone developmental research. Particularly, topics typically investigated in younger children are over-represented. Future analyses of a possible relation between age, effect size, and sample size would thus benefit from a larger sample of meta-analyses. A second potential impediment to generalizing from the presented findings to developmental research at large is the fact that we focused on language acquisition research. As there is no a priori reason to expect that sample sizes and effect sizes are particularly low in this sub-domain of developmental science, and because most methods are used across fields, we expect that the results and recommendations are relevant to researchers working in other domains. However, to be able to make such claims with more certainty, standardized collections of meta-analyses on phenomena in different sub-domains of developmental research are needed. We strongly encourage such endeavours, and have made all materials openly available and provided substantial documentation to expand this approach beyond language acquisition studies.

## Conclusion

We have showcased the use of standardized collections of meta-analyses for the diagnosis of (potential) issues in developmental research, using early language acquisition as a case study. Our results point to an overall lack of consideration of meta-analytic effect size in study planning, leading to habitually under-powered studies. In addition, method choice and participant age modulate effect size; we here provide first indicators of the importance of both factors in study design. To improve the replicability of developmental research, and as a consequence the empirical basis on which theories of development are built, we strongly

recommend an increased use of effect sizes and meta-analytic tools, including prospective power calculations.

## References

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

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101. doi:10.2307/2533446

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science, 19,* 901–917. doi:10.1111/desc.12341

Bergmann, C., Tsuji, S., & Cristia, A. (2017). Top-down versus bottom-up theories of phonological acquisition: A big data approach. *Proceedings of Interspeech 2017*, 2013–2016. doi:10.21437/Interspeech.2017-1443

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In *Proceedings of the 39th annual conference of the cognitive science society* (pp. 124–129). Cognitive Science Society.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. doi:10.1038/nrn3475

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

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. NJ: Lawrence Earlbaum Associates.

Colonnesi, C., Stams, G. J. J., Koster, I., & Noom, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review, 30*, 352–366. doi:10.1016/j.dr.2010.10.001

Cristia, A. (2017). Can infants learn phonology in the lab? A meta-analytic answer. In press at *Cognition*.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–Retest reliability in infant speech perception tasks. *Infancy*, *21*, 648–667. doi:10.1111/infa.12127

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177. doi:10.1037/1082-989X.1.2.170

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

Eason, A. E., Hamlin, J. K., & Sommerville, J. A. (2017). A survey of common practices in infancy research: Description of policies, consistency across and within labs, and suggestions for improvements. *Infancy*, *22*, 470–491. doi:10.1111/infa.12183

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy 22*, 421–435. doi:10.1111/infa.12182

Frank, M. C., Lewis, M., & MacDonald, K. (2016). A performance model for early word learning. In A. Papafragou, D. Grodner, Mirman D., & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2610–2614).

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development, 17*, 1–17. doi:10.1080/15248372.2015.1061528

Gredebäck, G., Johnson, S., & Hofsten, C. von. (2009). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*, 1–19. doi:10.1080/87565640903325758

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, *5*, 69–95.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124. doi:10.1371/journal.pmed.0020124

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*, 43–48. doi:10.1098/rspb.2001.1832

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*, 61–76. doi:10.1002/jrsm.35

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*, 278–292. doi:10.1177/1745691614528520

Lerner, R. M., Liben, L. S., & Mueller, U. (2015). *Handbook of child psychology and developmental science, cognitive processes* (Vol. 2). John Wiley & Sons.

Lewis, M. L., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., & Frank, M. C. (2017). A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis. *Preprint*. doi:10.17605/OSF.IO/HTSJM

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Sage publications Thousand Oaks, CA.

ManyBabies Collaborative. (2017). Quantifying sources of variability in infancy research using the infant-directed speech preference. Accepted pending data collection in *Advances in Methods and Practices in Psychological Science*.

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*, 416–432. doi:10.1111/infa.12078

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125. doi:10.1037/1082-989X.7.1.105

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14*, 1105–1107. doi:10.1038/nn.2886

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631. doi:10.1177/1745691612459058

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy, 22*, 436–469. doi:10.1111/infa.12186

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/

Shanks, D. R., Vadillo, M. A., Riedel, B., Clymo, A., Govind, S., Hickin, N., … Puhlmann, L. (2015). Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? *Journal of Experimental Psychology: General, 144,* e142–e158. doi:10.1037/xge0000116

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3,* 160384. doi:10.1098/rsos.160384

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49,* 108–112. doi:10.1080/00031305.1995.10476125

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology, 56,* 179–191. doi:10.1002/dev.21179

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Psychological Science, 9,* 661–665. doi:10.1177/1745691614552498

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23,* 87–102. doi:10.3758/s13423-015-0892-6

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

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science, 4,* 294–298. doi:10.1111/j.1745-6924.2009.01127.x

## Tables

Table 1. *Descriptions of the meta-analyses. Age is reported in months, sample size is based on the median in a given meta-analysis, effect size is reported as meta-analytic weighted median Cohen's d, and average power is computed based on meta-analytic effect size estimate Cohen's d and median sample size.*

| Meta-Analysis | Age | Sample Size | N Effect Sizes | N Papers | Effect Size (SE) | Power |
|---|---|---|---|---|---|---|
| Gaze following | 14 (3-24) | 23 (12-63) | 32 | 11 | 1.08 (0.16) | 0.95 |
| IDS preference | 4 (0-9) | 20 (10-60) | 48 | 16 | 0.73 (0.13) | 0.61 |
| Concept-label advantage | 12 (4-18) | 13 (9-32) | 48 | 15 | 0.45 (0.08) | 0.20 |
| Mutual exclusivity | 24 (15-60) | 16 (8-72) | 58 | 19 | 0.81 (0.14) | 0.61 |
| Online word recognition | 18 (15-30) | 25 (16-95) | 14 | 6 | 1.24 (0.26) | 0.99 |
| Phonotactic learning | 11 (4-16) | 18 (8-40) | 47 | 15 | 0.12 (0.07) | 0.06 |
| Pointing and vocabulary | 22 (9-34) | 24.5 (6-50) | 12 | 12 | 0.98 (0.18) | 0.92 |
| Sound symbolism | 8 (4-38) | 20 (11-40) | 44 | 11 | 0.22 (0.11) | 0.10 |
| Statistical sound learning | 8 (2-11) | 15.5 (5-34) | 19 | 11 | 0.29 (0.14) | 0.12 |
| Native vowel discrimination | 7 (0-30) | 12 (6-50) | 112 | 29 | 0.69 (0.09) | 0.37 |
| Non-native vowel discrimination | 8 (2-18) | 16 (8-30) | 46 | 14 | 0.79 (0.24) | 0.58 |
| Word segmentation | 8 (6-25) | 20 (4-64) | 284 | 68 | 0.16 (0.03) | 0.08 |

Table 2. *For each meta-analysis, largest effect size Cohen's d and derived power based on the seminal paper, along with the difference between power based on meta-analytic and seminal paper effect size.*

| **Meta-Analysis** | **Effect Size (Seminal)** | **Effect Size (Overall)** | **Sample Size** | **Power (Seminal)** | **Difference** |
|---|---|---|---|---|---|
| Statistical sound learning | -0.24 | 0.29 | 15.5 | 0.10 | -0.02 |
| Word segmentation | 0.56 | 0.16 | 20 | 0.40 | 0.33 |
| Mutual exclusivity | 0.70 | 0.81 | 16 | 0.48 | -0.13 |
| Concept-label advantage | 0.86 | 0.45 | 13 | 0.56 | 0.36 |
| Pointing and vocabulary | 0.65 | 0.98 | 24 | 0.61 | -0.31 |
| Non-native vowel discrimination | 1.02 | 0.79 | 16 | 0.80 | 0.22 |
| Phonotactic learning | 0.98 | 0.12 | 18 | 0.81 | 0.75 |
| Sound symbolism | 0.95 | 0.22 | 20 | 0.84 | 0.73 |
| Online word recognition | 0.89 | 1.24 | 25 | 0.87 | -0.12 |
| Gaze following | 1.29 | 1.08 | 23 | 0.99 | 0.04 |
| Native vowel discrimination | 1.87 | 0.69 | 12 | 0.99 | 0.63 |
| IDS preference | 2.39 | 0.73 | 20 | 1.00 | 0.39 |

Table 3. *Linear mixed effects model predicting exclusion rate by method and participant age while accounting for the specific phenomenon, central fixation is the baseline method. CondHT = conditioned headturn, FC = forced choice, HPP = headturn preference procedure, LwL = looking while listening, SA = stimulus alternation.*

| | **Est.** | **SE Est** | **t** | **p** |
|---|---|---|---|---|
| Intercept | 31.170 | 4.481 | 6.96 | **<.001** |
| CondHT | 31.064 | 5.727 | 5.42 | **<.001** |
| FC | -26.383 | 9.372 | -2.82 | **.005** |
| HPP | -2.132 | 4.770 | -0.45 | .655 |
| LwL | -6.433 | 5.394 | -1.19 | .233 |
| SA | 21.345 | 4.129 | 5.17 | **<.001** |
| Age | 0.409 | 0.438 | 0.93 | .350 |
| CondHT*Age | 2.888 | 1.160 | 2.49 | **.013** |
| FC*Age | -0.207 | 0.645 | -0.32 | .749 |
| HPP*Age | 0.975 | 0.717 | 1.36 | .174 |
| LwL*Age | -0.548 | 0.796 | -0.69 | .491 |
| SA*Age | -0.251 | 0.903 | -0.28 | .781 |

Table 4. *Meta-analytic regression predicting effect size Cohen's d with participant age and method (central fixation is baseline method). CondHT = conditioned headturn, FC = forced choice, HPP = headturn preference procedure, LwL = looking while listening, SA = stimulus alternation.*

| | **Est. (CI)** | **SE** | **z** | **p** |
|---|---|---|---|---|
| Intercept | 0.285 [0.005,0.566] | 0.143 | 2.00 | **.046** |
| Age | 0.014 [-0.002,0.026] | 0.006 | 2.25 | **.024** |
| CondHT | 1.284 [0.627,1.94] | 0.335 | 3.83 | **<.001** |
| FC | 0.109 [-0.261,0.48] | 0.189 | 0.58 | .563 |
| HPP | 0.125 [-0.043,0.293] | 0.086 | 1.46 | .144 |
| LwL | 0.498 [0.071,0.925] | 0.218 | 2.29 | **.022** |
| SA | -0.141 [-0.506,0.224] | 0.186 | -0.76 | .449 |
| Age*CondHT | 0.107 [-0.003,0.217] | 0.056 | 1.91 | .056 |
| Age*FC | 0.044 [0.028,0.059] | 0.008 | 5.51 | **<.001** |
| Age*HPP | 0.006 [-0.013,0.024] | 0.010 | 0.60 | .546 |
| Age*LwL | 0.019 [-0.002,0.041] | 0.011 | 1.80 | .071 |
| Age*SA | -0.005 [-0.057,0.047] | 0.027 | -0.02 | .845 |

Table 5. *Non-parametric correlations between sample sizes and effect sizes for each meta-analysis. A significant value indicates bias.*

| **Meta-analysis** | **Kendall's Tau** | **p** |
|---|---|---|
| Phonotactic learning | -0.21 | .052 |
| Statistical sound learning | -0.06 | .724 |
| Gaze following | 0.09 | .512 |
| IDS preference | 0.01 | .921 |
| Concept-label advantage | -0.06 | .590 |
| Mutual exclusivity | -0.21 | **.024** |
| Native vowel discrim. | -0.28 | **<.001** |
| Non-native vowel discrim. | -0.23 | **.032** |
| Pointing and vocabulary | -0.15 | .491 |
| Sound symbolism | -0.04 | .698 |
| Online word recognition | -0.13 | .539 |
| Word segmentation | -0.10 | **.023** |

**Figures**

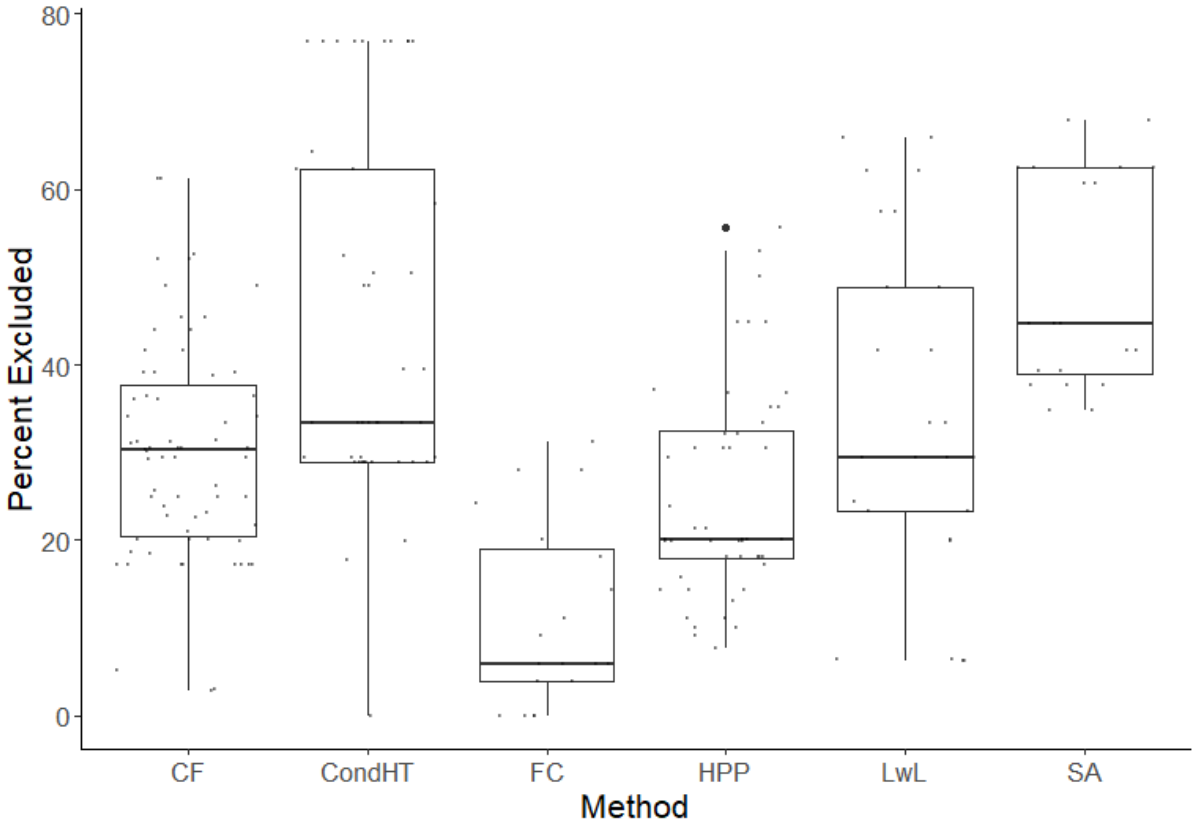


Figure 1. *Exclusion rate in percent by different methods. CF = central fixation, CondHT = conditioned headturn, FC = forced choice, HPP = headturn preference procedure, LwL = looking while listening, SA = stimulus alternation. Each point indicates a single study.*

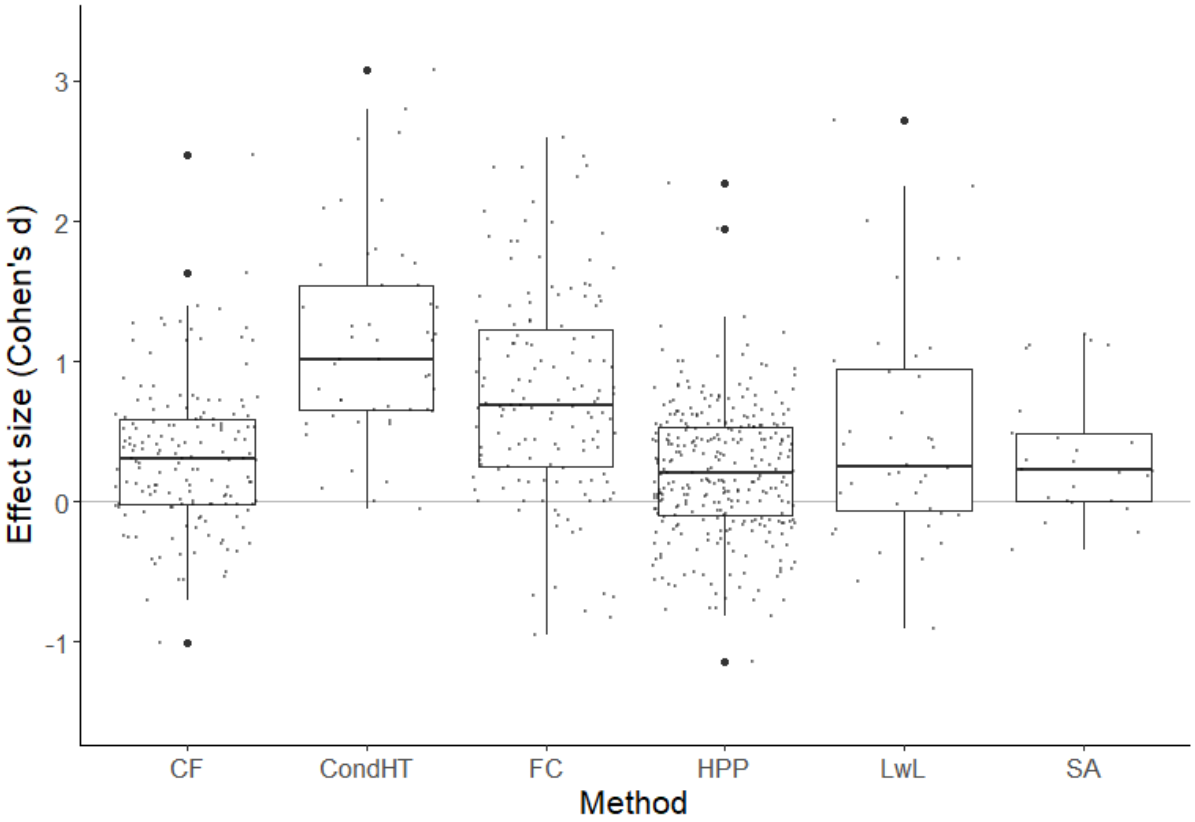


Figure 2. *Effect size by different methods. CF = central fixation, CondHT = conditioned headturn, FC = forced choice, HPP = headturn preference procedure, LwL = looking while listening, SA = stimulus alternation. Each point indicates a single study.*

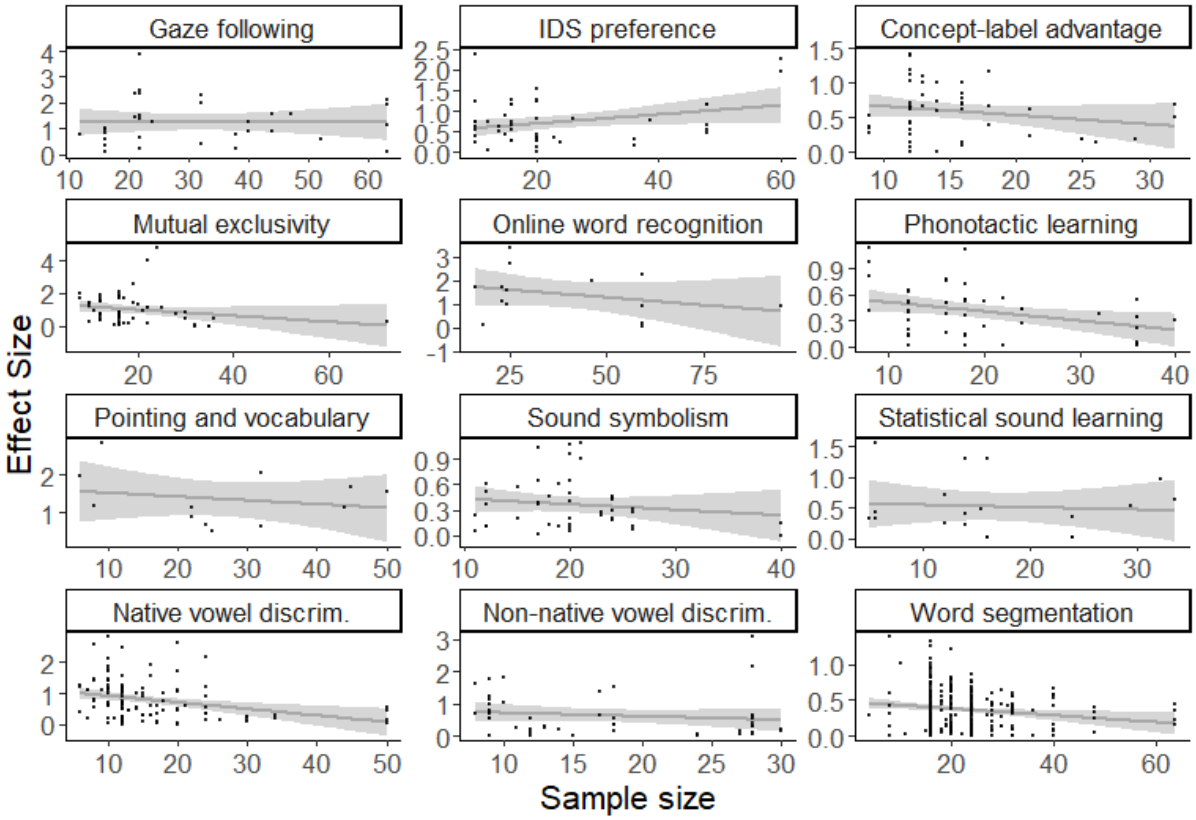


Figure 3. *For every meta-analysis observed effect size per study plotted against sample size. Each point indicates a single study.*

**Acknowledgements**

The authors of this paper thank all contributors and supporters of MetaLab as well as the anonymous reviewers whose thoughtful comments helped improve the paper.