Assessing methodological practices in language acquisition research through meta-analyses

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

[1] Ecole Normale Sup{'e}rieure, PSL Research University, D{'e}partement d'Etudes Cognitives, Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)

[2] University of Pennsylvania, Department of Psychology

[3] Ecole Normale Sup{'e}rieure, PSL Research University, D{'e}partement d'Etudes Cognitives, Neuropsychologie Interventionnelle (ENS, EHESS, CNRS)

[4] University of Chicago, Computation Institute/University of Wisconsin-Madison, Department of Psychology

[5] Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences

[6] Stanford University, Department of Psychology, Language and Cognition Lab

Author Note

Correspondence concerning this article should be addressed to Christina Bergmann, Ecole Normale Su{'e}rieure, Laboratoire de Sciences Cognitives et Psycholinguistique, 29, rue d'Ulm, 75005 Paris, France.. E-mail: chbergma@gmail.com

Abstract

Replicability is a critical feature of scientific research. Recent concerns about replicability in psychology have led to a focus on statistical power, the probability of experiments to detect particular effects. Using data from a new database of meta-analyses, we analyze methodological trends in the language development literature. Although statistical power has been a concern in infancy research, no extant data speak to the average level of power in this area. We calculated the typical statistical power for experiments across our database by comparing sample sizes in each experiment to the meta-analytic estimate of the effect size. With a median effect size of Cohen's $d = .57$ across all 12 phenomena, and a typical sample size of 17 participants per cell, power is at 60%. This finding suggests that typical sample sizes in infancy research are likely too low and that researchers do not habitually consider effect sizes in their experiment planning. We also show that seminal publications in the early language learning literature typically over-estimate effect sizes relative to later investigations, but that this literature does not show evidence of "p-hacking" (undisclosed analytic flexibility). We conclude with recommendations for experimental planning and reporting as well as for the use of meta-analysis in developmental research.

Keywords: replicability, reproducibility, meta-analysis, language acquisition, power

Word count: X

Assessing methodological practices in language acquisition research through meta-analyses

Empirical research is built on a never-ending conversation between theory and data, between expectations and observations. Theories lead to new experimental questions and new data in turn help us refine our theories. This process relies crucially 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 in turn often depends on the observation of statistically significant and theoretically-surprising outcomes. If researchers aim for publishability, this is likely to lead to practices that undermine the quality and reliability of their data. It has therefore been suggested that theories should rely on replicable findings. Replicability is crucial in experimental sciences, particularly for developmental research: Theories should be based on robust findings and their boundary conditions have to be explored with sufficiently powered studies to avoid an excess of false negatives. Further, translating findings on child development into practice requires a solid knowledge base.

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). Simulating the scientific process, Ioannidis (2005) speculated that most empirical research findings may even be false. The proportion of false findings in these simulations was dependent on several features, including the underlying effect size of a particular phenomenon, the typical sample sizes used by researchers, and the degree of flexibility in data collection and analysis. All of these factors are highly relevant to developmental research.

In the current paper, we survey and quantify methodological practices in developmental research using meta-analytic tools, focusing on language development. We take a different approach from the typical meta-analysis by aggregating over multiple datasets. Using a collection of standardized meta-analyses, we focus on key experimental design choices: sample size (and the ensuing statistical power) and experimental method. In doing so, we provide what is, to our knowledge, the first assessment of typical practices of developmental research. Based on our findings and experiences with building meta-analyses and using meta-analytic tools, we end this paper with suggestions for change.

The data we analyze are part of MetaLab, a database of meta-analyses of language acquisition that, covers a variety of methods (11 in total) and participant ages, from newborns to 3.50-year-olds. Since our approach is accompanied by extensive educational materials, completely open data and scripts, and we build on open source software (particularly R; R Core Team, 2016), our approach can easily be extended to other domains of child development research and we strongly encourage fellow researchers to build similar collections of meta-analyses describing and quantifying phenomena in their sub-domain of developmental research.

Key concerns for robust research in developmental science

In this section we review potential hindrances to developmental research being robust and reproducible, and briefly describe how we will assess the status quo. Note that all these descriptions are by necessity brief, for extended discussions we provide references to suitable readings.

Statistical power. 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 therefore dependent on the underlying effect size and the sample size. Of course, low power is problematic in terms of increased chances of type-II errors (i.e., failure to find a significant result when there is an underlying effect). It has become increasingly clear that low power is also problematic in the case of type-I errors, or false positives, as the effects reported in such cases will be over-estimating the true effect (Button et al., 2013; see also Ioannidis, 2005; Simmons et al., 2011). This makes appropriate planning for future research more difficult, as sample sizes will be too small, leading to null results due to insensitive research designs rather than the absence of the underlying effect. This poses a serious hindrance for 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 non-significant 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 (**???**) by using data from a high-powered looking time study.

To investigate the status quo, we first compute typical power per phenomenon, based on meta-analytic effect sizes and typical sample size. We explore which effect sizes would be detectable with the sample sizes present in our datasets. We additionally investigate how researchers might determine sample sizes using a different heuristic, following the first paper on their phenomenon of interest.

Method choice. Improving procedures in developmental research can be considered both an economical and ethical necessity, because the population is difficult to recruit and test. For this reason, developmentalists often "tweak" paradigms and develop new ones to increase reliability and robustness, all with the aim of obtaining a clearer signal. Especially given the time constraints, we aim to collect a maximum of data in the short time span infants and children are willing to participate in a study. Emerging technologies, such as eye-tracking and tablets, have consequently been eagerly adopted (Frank, Sugarman, Horowitz, Lewis, and Yurovsky, 2016). As a result, multiple ways to tap into the same phenomenon exist; consider for example the fact that both headturn-based paradigms and offline as well as online measurements of eye movements are frequently being employed to measure infant-directed speech preference (Dunst, Gorman, & Hamby, 2012; ManyBabies Collaborative, 2017).

It remains an open question to what extent these different methods lead to comparable results. It is possible that some are more robust, but it is difficult to extract such information based on single studies that use different materials and test various age groups (but see the large-scale experimental approach by ManyBabies Collaborative, 2017). Aggregating over experimental results via meta-analytic tools, in contrast, allows us to extract general patterns of higher or lower noise by comparison of effect sizes, which are directly affected by the variance of the measurement.

We will assess in how far the different methods used in the present collection of meta-analyses vary in the resulting effect size. Further, taking possible resource limitations into account, we consider drop-out rates as a potential measure of interest and discuss whether higher exclusion rates coincide with more precise measures, yielding higher effect sizes.

Questionable research practices. Undisclosed flexibility during data collection and analysis is a problem independent of the availability of various methods to conduct developmental studies. To begin with, using flexible stopping rules, where the decision to stop or continue testing depends on the result of a statistical test, increases the likelihood to obtain a "significant" outcome well beyond the traditional 5%.

As for analytic flexibility, researchers might conduct multiple significance tests with several more or less related dependent variables without correcting for this practice. In developmental research, this encompasses transforming the same measured data into multiple dependent variables (such as mean scores, difference scores, percentages, and so on) as well as selectively excluding trials and re-testing the new data for statistical significance. Next, multiple conditions that selectively can be dropped from the final report increase the number of significance tests. Finally, it is problematic to post hoc introduce covariates, most prominently gender, and test for an interaction with the main effect, and solely report those outcomes as confirmatory hypothesis test. Combining two or more of these strategies again increase the number of significant results that occur by chance even if there is no effect present in the population. All these practices might seem innocuous and geared towards "bringing out" an effect the researcher believes is real, yet they can inflate the number of significant p values, 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 flexibility led to a false positive in a given report. However, we can measure "symptoms" of such practices in a whole literature. We focus in this paper on flexibility in stopping, 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 (publications date range from 1973 to 2017), it might be the case that recent discussions of best practices have improved lab practices, but older reports could still have applied this seemingly innocuous practice of adding participants to "bring out" the effect of interest.

Methods

Data

All data presented and analyzed in the present paper are part of a standardized collection of meta-analyses (MetaLab), and are freely available via the companion website http://metalab.stanford.edu. Currently, MetaLab contains 13 meta-analyses, or datasets, 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, native language), methodological information (for example what dependent variable was measured), and information necessary to compute effect sizes (number of participants, if available means and standard deviations of the dependent measure, otherwise test statistics of the key hypothesis test, such as t values or F scores). This way, the analyses presented in this paper become possible.

MetaLab contains datasets that address phenomena ranging from infant-directed speech preference to mutual exclusivity, sampled opportunistically. Meta-analyses are either based on data made available on MetaLab by their original authors (n=11 datasets) 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 original 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 dependent measure, we report Cohen's *d*, a standardized effect size based on comparing sample means and their variance. Effect size was calculated when possible from means and standard deviations across designs with the appropriate formulae. When these data were not available, we used test statistics, more precisely t values or F scores of the test assessing the main hypothesis. We also computed effect size variance, which allows to weight 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 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. Some correlations could be obtained through direct contact with the original authors (see e.g., Bergmann & Cristia, 2016 for details), 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.

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). We relied on the implementation in the metafor package (Viechtbauer, 2010) of R (R Core Team, 2016). Excluded as outliers were effect sizes more than three standard deviations away from the median effect size within each dataset, thus accounting for the difference in median effect size across phenomena.

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 phenomenon. This approach is insightful, because meta-analytic effect size estimates are (typically) more reliable than those of single studies. For targeted analyses of the power of the seminal paper, 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.

Results

Statistical power

Table 1 provides a summary of typical sample sizes and effect sizes by phenomenon. We remind the reader that recommendations are for this value to be above 80%, which refers to a likelihood that four out of five studies show a significant outcome for an effect truly present in the population.

As could be expected, sample sizes are small across all phenomena, with the overall median in our data being 17. Effect sizes tend to fall into ranges of small to medium effects, as defined by Cohen (Cohen, 1988). The overall median effect size of all datasets is Cohen's $d = 0.69$. 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) it is possible to detect an effect in 80% of all studies when Cohen's $d = 0.72$; 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 17 participants per group increases to Cohen's $d = 0.99$, 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 $d = 0.69$, studies would have to test 18 participants in a paired design; 1 more than are included on average. It should be noted that this disparity between observed and necessary sample size varies greatly across phenomena, leading to drastic differences in observed power to detect the main effect at stake. While studies on phonotactic learning and word segmentation apparently typically run dramatically underpowered studies (with typical power being under 10%), experiments on gaze following and online word recognition are very highly powered (95% and 99%, respectively).

NOTE: Will ad summary row in the bottom by hand

The role of participant age. Participant age can be assumed to interact with effect size both for conceptual and practical reasons. Younger participants might show smaller effects in general because they are more immature in terms of their information processing abilities, and they are not yet as experienced with, and proficient in, their native language in particular. As to practical reasons, measurements might be more noisy for younger participants, as they could be a more difficult population to recruit and test. We find no linear relationship between participant age and sample size, effect size, and derived power on the level of meta-analyses. In addition, the prediction that older participants might be easier to recruit and test is not reflected in the observed sample sizes. However, the only two datasets, gaze following and online word recognition, with power over 80% typically test participants older than one year.

Seminal papers as basis for sample size planning. As Table 1 shows, experimenters are frequently not including a sufficient number of participants to observe a given effect – assuming the meta-analytic estimate is accurate. It might, however, be

Table 1

Descriptions of meta-analyses.

Topic	Age (in Months)	Median Sample Size (Range)	N Effect Sizes
Infant directed speech preference	4.34	20 (10, 60)	48
Vowel discrimination (native)	6.54	12 (6, 50)	112
Vowel discrimination (non-native)	7.69	16 (8, 30)	46
Sound symbolism	7.89	20 (11, 40)	44
Statistical sound category learning	8.16	14.75 (5, 35)	16
Word segmentation	8.29	20 (4, 64)	284
Phonotactic learning	10.69	18 (8, 40)	47
Label advantage in concept learning	12.36	13 (9, 32)	48
Gaze following	14.24	23 (12, 63)	32
Online word recognition	18.00	25 (16, 95)	14
Pointing and vocabulary (concurrent)	22.22	24.5 (6, 50)	12
Mutual exclusivity	23.99	16 (8, 72)	58
Categorization Bias	42.00	14 (8, 20.5)	77

possible, that power has been determined based on a seminal paper to be replicated and/or expanded. Initial reports tend to overestimate effect sizes (Jennions & Møller, 2002), possibly explaining the lack of power in some datasets and studies.

We extracted for each dataset the oldest paper and therein the largest reported effect size and re-calculated power accordingly, using the median sample size of a given dataset. The results are shown in Table 2. It turns out that in some cases, such as native and non-native vowel discrimination, sample size choices match well with the oldest report. The difference in power, noted in the last column, can be substantial, with native vowel discrimination and phonotactic learning being the two most salient examples. Here, sample sizes match well with the oldest report and studies would be appropriately powered if this estimate were representative of the true effect. For four datasets neither the seminal paper nor meta-analytic effect size seem to be basis for sample size decisions.

Method choice

In most of our meta-analyses, multiple methods were used to tap into the phenomenon at stake. Choosing a robust method can help increase power, because 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 participants had to be invited into the lab. We thus first quantify whether methods differ in their typical drop-out rate, as economic considerations might drive method choice. To this end we consider all methods across datasets which have more than 10 associated effect sizes and for which information on the number of dropouts was reported; this information is not always reported in published papers. In the case of the two meta-analyses we added based on published reports, the information of drop-out rates was not available. Therefore, the following analyses only cover 6 methods and 224 data points.

Drop-out rates across procedures. The results of a linear mixed effects model predicting dropout rate by method and mean participant age (while controlling for the different phenomena and associated underlying effect sizes being tested) are summarized in the table below. The results show that, taking the most frequently used method central fixation as the baseline, conditioned headturn and stimulus alternation have significantly more drop-outs, while forced choice has significantly fewer. Figure 1 underlines this observation. Overall, stimulus alternation leads to the highest drop-out rates, which lies at around 50% (see Figure 1), and forced choice to the lowest. Participant age interacts with the different methods. We observe an increase in drop-out rates, which is most prominent in conditioned headturn (a significant interaction) and headturn preference procedure (where

Table 2

*For each meta-analysis, largest effect size Cohen's *d* and derived power based on the first paper, along w*

oldest effect size.

Meta-analysis (MA)	Oldest Effect Size	Meta-analytic Effect Size	Sample Size	Po
Statistical sound category learning	0.56	-0.26	15	
Word segmentation	0.56	0.16	20	
Mutual exclusivity	0.70	0.81	16	
Label advantage in concept learning	0.86	0.45	13	
Pointing and vocabulary (concurrent)	0.65	0.98	24	
Vowel discrimination (non-native)	1.02	0.79	16	
Phonotactic learning	0.98	0.12	18	
Sound symbolism	0.95	0.22	20	
Online word recognition	0.89	1.24	25	
Gaze following	1.29	1.08	23	
Vowel discrimination (native)	1.87	0.69	12	
Infant directed speech preference	2.39	0.73	20	
Categorization Bias	9.06	0.27	14	

Table 3

Linear mixed effects model predicting dropout rate by method and participant age while accounting for the s
phenomenon.

	Estimate	Std. Error	t value
(Intercept)	32.8372364547405	5.18270575595169	6.33592528710
methodconditioned head-turn	41.7568152334524	9.70666233683515	4.30187162017
methodforced-choice	-27.2548832121313	8.87560255001973	-3.0707642730
methodhead-turn preference procedure	1.40027320022283	6.35607050157499	0.22030485657
methodlooking while listening	-8.61485990923845	6.92980597857125	-1.2431603331
methodstimulus alternation	20.3310364668288	6.33917912667968	3.20720333982
ageC	0.419580203901804	0.439671895960652	0.95430298765
methodconditioned head-turn:ageC	2.87744919382107	1.16473791428906	2.47046924335
methodforced-choice:ageC	-0.2158939800302	0.647747533970996	-0.3332995784
methodhead-turn preference procedure:ageC	0.963028502288058	0.719580640550116	1.33831908200
methodlooking while listening:ageC	-0.567347683657796	0.798508880012725	-0.7105089221
methodstimulus alternation:ageC	-0.261530192918133	0.907336467333541	-0.2882394815

the interaction approaches significance).

Interestingly, the methods with lower drop-out rates, namely central fixation and headturn preference procedure, are among the most frequent ones in our data and certainly more frequent than those with higher drop-out rates. The proportion of participants that can be retained might thus indeed 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.

Methods which retain a higher percentage of participants might either be more

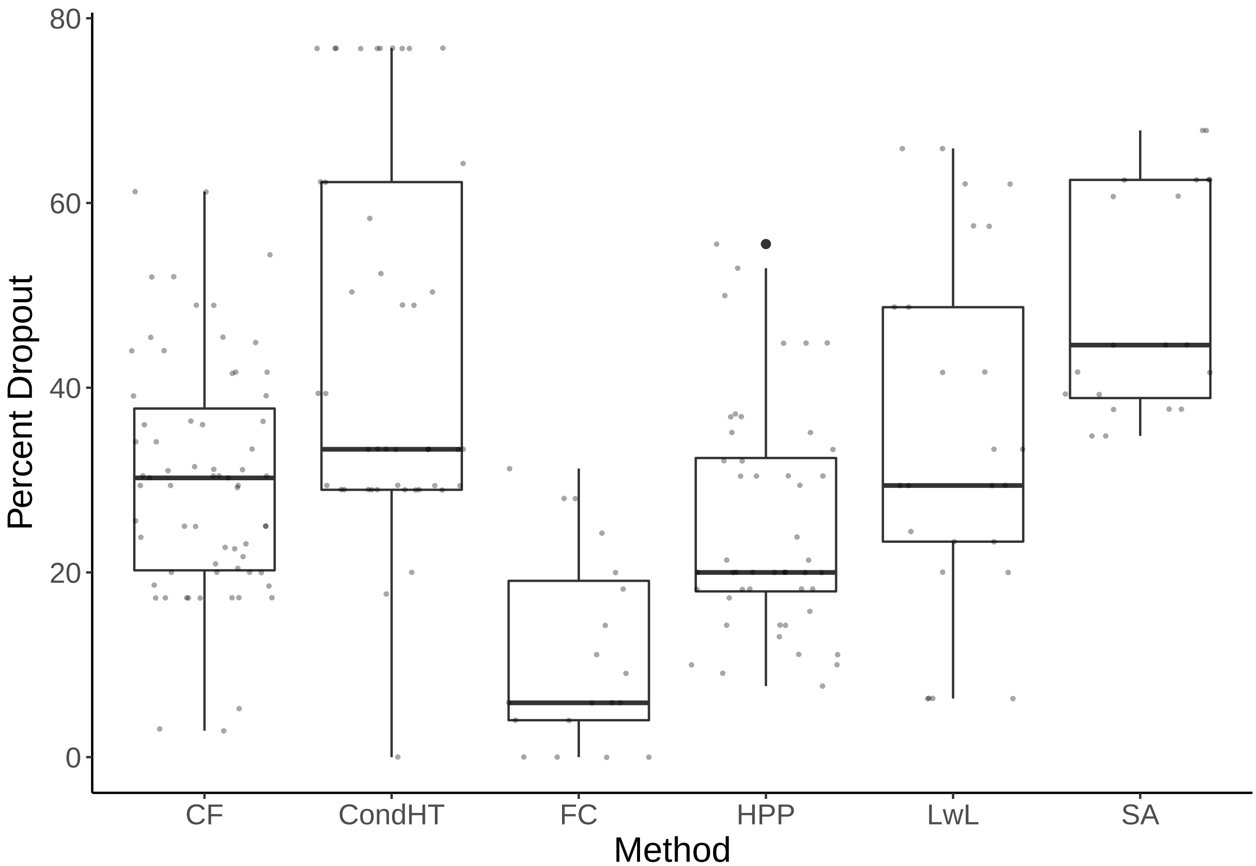

Figure 1. Percent dropout as explained by different methods. CF = central fixation, CondHT = conditioned headturn, FC = forced choice, HPP = headturn preference procedure, LwL = looking while listening, SA = stimulus alternation.

suitable, because they are decreasing noise as most participants are on task, or less selective, thus increasing noise as participants who for example are fussy are more likely to enter the data pool. We thus turn to a meta-analytic assessment of the same methods discussed here.

Effect sizes as a function of procedure. We built a meta-analytic model with Cohen's d as the dependent variable, method and mean age centered as independent variables, which we allowed to interact. The model includes the variance of d for sampling variance, and paper within meta-analysis as a random effect nested within phenomenon (because we assume that within a paper experiments and thus effect sizes will be more similar to each other than across papers). We again selected the most frequently used

Table 4

*Meta-analytic regression predicting effect size Cohen's *d* with participant age and method (central fixatio*

	estimate	se	zval	pval
intrcpt	0.2240	0.1440	1.56	0.119
ageC	0.0112	0.0064	1.75	0.079
relevel(method, "central fixation")conditioned head-turn	1.8230	0.6047	3.01	0.002
relevel(method, "central fixation")forced-choice	0.5223	0.1868	2.80	0.005
relevel(method, "central fixation")head-turn preference procedure	0.1832	0.1163	1.58	0.115
relevel(method, "central fixation")looking while listening	0.4402	0.2427	1.81	0.069
relevel(method, "central fixation")stimulus alternation	-0.0626	0.2745	-0.23	0.819
ageC:relevel(method, "central fixation")conditioned head-turn	0.1144	0.0629	1.82	0.069
ageC:relevel(method, "central fixation")forced-choice	-0.0089	0.0065	-1.36	0.172
ageC:relevel(method, "central fixation")head-turn preference procedure	0.0091	0.0098	0.94	0.349
ageC:relevel(method, "central fixation")looking while listening	0.0250	0.0111	2.24	0.024
ageC:relevel(method, "central fixation")stimulus alternation	0.0039	0.0281	0.14	0.889

method central fixation as the baseline and limited this analysis to the same methods that we investigated above.

The model results in Table 2 show that compared to central fixation, conditioned headturn and forced choice yield reliably higher effect sizes, all other methods do not statistically differ from this baseline (note that looking while listening is approaching significance). When factoring in age, looking while listening shows a significant interaction, and conditioned headturn approaches significance, indicating an increase in effect sizes as infants mature. Age is marginally above the significance threshold, the positive estimate further underlines that overall effect sizes increase for older participants – an observation

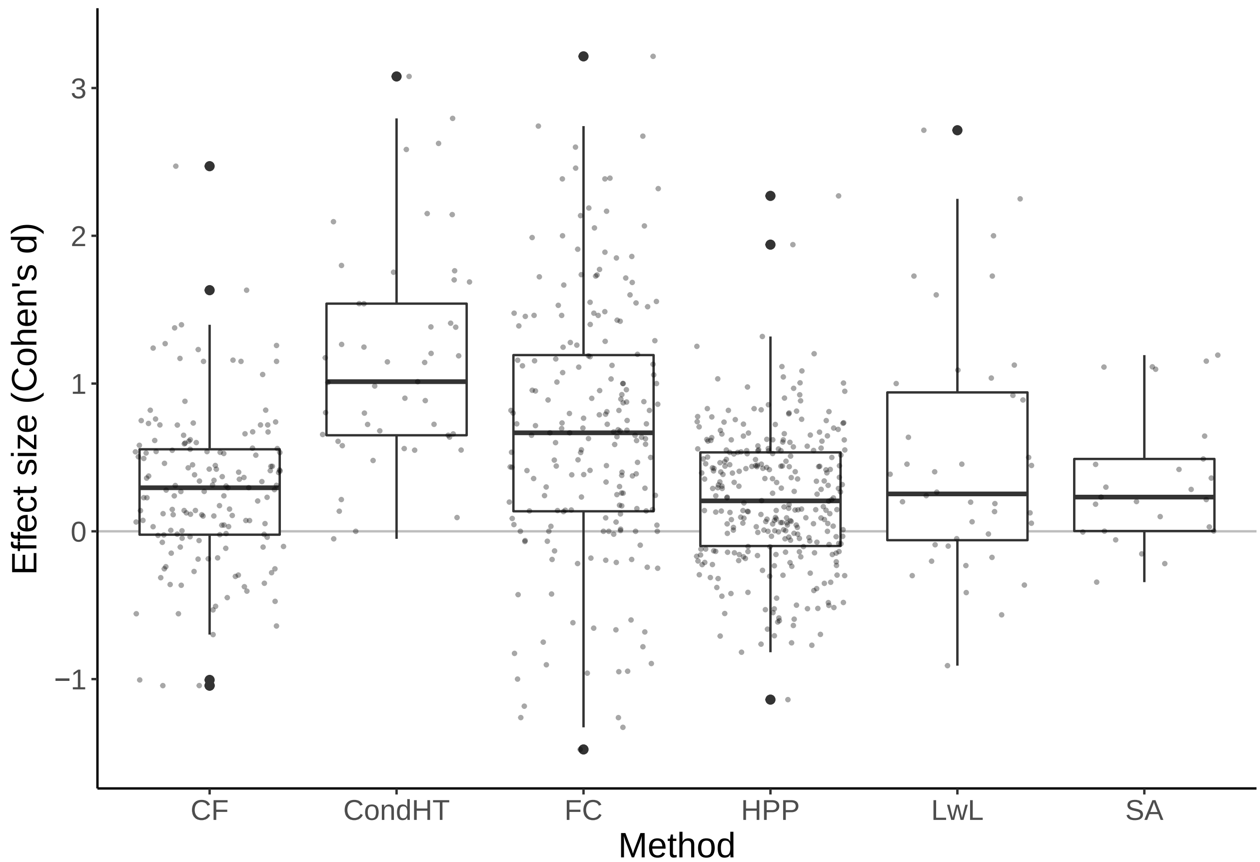

Figure 2. Effect size by different methods. CF = central fixation, CondHT = conditioned headturn, FC = forced choice, HPP = headturn preference procedure, LwL = looking while listening, SA = stimulus alternation.

consistent with the view that infants and toddlers become more proficient language users and are increasingly able to react appropriately in the lab.

Questionable research practices

To assess whether researchers selectively add participants to obtain a significant p value, we assess the relationship 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, 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. A significant non-parametric correlation indicates that only those studies with significant outcomes were published (Begg & Mazumdar, 1994).

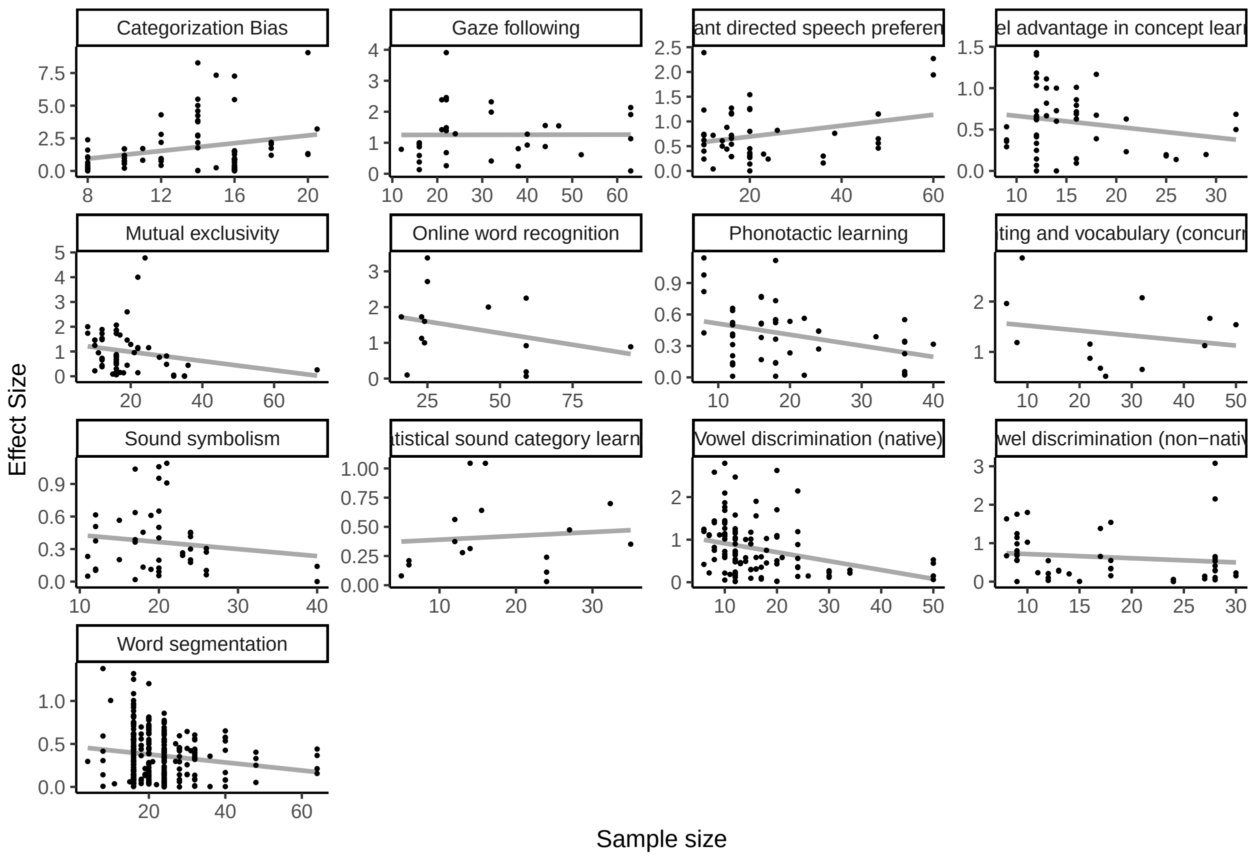

Figure 3. For each dataset observed effect size per study plotted against sample size.

We illustrate the relationship between effect size and sample size, separated by meta-analysis, in Figure XX. The regression line is plotted on top of points indicating single experiments. Four datasets turn out to have a significant negative relationship between sample size and effect size, indicating bias; two assessing infants' ability to discriminate vowels, one on word segmentation, and one testing whether children use mutual exclusivity during word learning. The last case might be driven by a single high-powered study, however. We further observe a positive relationship between sample size and observed effect size in two datasets, namely infant directed speech preference and categorization bias.

Table 5

Non-parametric correlations between sample sizes and effect sizes for each dataset. A significant value indicates bias.

Meta-analysis	Kendall's Tau	p-value
Phonotactic learning	-0.207	0.052
Statistical sound category learning	0.205	0.277
Categorization Bias	0.151	0.07
Gaze following	0.085	0.512
Infant directed speech preference	0.010	0.921
Label advantage in concept learning	-0.057	0.59
Mutual exclusivity	-0.214	0.024
Vowel discrimination (native)	-0.283	$< .001$
Vowel discrimination (non-native)	-0.229	0.032
Pointing and vocabulary (concurrent)	-0.154	0.491
Sound symbolism	-0.042	0.698
Online word recognition	-0.128	0.539
Word segmentation	-0.098	0.023

Discussion

In this paper, we made use of a collection of standardized meta-analyses to assess the status quo in developmental research regarding typical effect sizes, sample size, power, and methodological choices in 12 meta-analyses on language development. With an average meta-analytic effect size of .57 and a typical sample size of only 17 participants per cell, we find that power is at 60%.

This means studies on language development, the sub-domain of developmental research that the present collection of meta-analyses is focused on, are severely under-powered. This 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. Phenomena targeting older children tend towards larger effects, and here some studies turn out to be high-powered (see for example online word recognition). Both observations are first indicators that effect size estimates might not be considered when determining sample size. It might, in the case of apparently over-powered studies however be possible that next to testing a main effect, such as whether children recognize a given word online, studies aimed to tap into factors affecting this ability. In this case, studies would be powered appropriately, as an interaction effect will be more difficult to detect than a main effect. Nonetheless, by and large, we find that studies are habitually underpowered

We investigated the alternative possibility that researchers base their sample size on the effect size reported in the seminal paper of their research topic. This turns out to be an unsuitable strategy: As described in the results section, the larger the original effect size, the more likely is an overestimation of the meta-analytic effect size. Researchers might thus be wary of reports implying a strong, robust effect with infants and toddlers in the absence of corroborating data. The lack of a relationship between either overall meta-analytic effect size or seminal reported effect size and sample size across phenomena indicates that researchers' experiment planning is not impacted by an estimated effect size of the phenomenon under investigation. Studies might instead be designed and conducted with pragmatic considerations in mind, such as participant availability.

To help researchers choose the most efficient experiment design, and thus potentially improve their power due to the use of a more sensitive measure, we next turned to methods. Our investigation of method choice considered both drop-out rates and whether effect sizes are differing across methods. Overall, drop-out rates varied a great deal (with medians between 5.9% for forced-choice and 45% for stimulus alternation). However, high drop-out rates might be offset by high effect sizes – at least in the case of conditioned headturn. While drop-out rates are around 30-50%, effect sizes are above 1. Stimulus alternation, in contrast, does not fall into this pattern of high drop-out rates being correlated with high effect sizes, as the observed effect sizes associated with this method are in the range typical for meta-analyses in our dataset. The interpretation of this finding might be that some methods, specifically conditioned headturn, which have higher dropout rates, are better at generating high effect sizes due to decreased noise (e.g., by excluding participants that are not on task). However, there is an important caveat: Studies with fewer participants (thanks to higher drop-out rates) might simply be underpowered, and thus any significant finding is likely to over-estimate the effect. We conclude thus that current efforts to estimate the impact of method choice experimentally are an important endeavor in developmental research (Frank et al., 2016).

A final set of analyses assessed the relationship between observed effect size and sample size. This analysis might reflect whether researchers selectively add participants to obtain a significant result. We observed that in four datasets smaller effect sizes coincided with larger sample sizes, which might be an indication of questionable research practices. At the same time we find two (numerically) positive correlations, an unexpected result as it means that larger sample sizes coincide with larger effects. One possible reason for this might be that for example older infants are both easier to test and yield larger effects. This explanation is in line with our finding when investigating the effect of method that higher participant age is linked to larger effect sizes.

For the observed negative correlations alternative explanations to questionable research practices are also possible: As soon as researchers are aware that they are measuring a more subtle effect (for example when selecting a contrast that is acoustically more difficult to distinguish or when testing younger infants) and adjust sample sizes accordingly, we expect to observe this negative correlation. In fact, in the presence of consequent and accurate a priori power calculations, a 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.

We have assessed the same datasets for other indicators of questionable research practices and publication bias, namely p-curves (**???**) and funnel plot asymmetry. For three datasets that showed a negative correlation between sample size and effect size, we also observe funnel plot asymmetry (both datasets on vowel discrimination as well as mutual exclusivity). For those three datasets we can thus conclude that publication bias underlies the observed link between sample size and effect size.

Concrete recommendations for developmental scientists

In this section, we aim to show how to move on from the status quo and improve the reliability of developmental research.

1. Calculate power prospectively. Our results indicate that most studies testing infants and toddlers are severely underpowered when aiming to detect a main effect. Interactions will show smaller effect sizes and thus will be even harder to detect in most cases. Further, power varies greatly across phenomena, which mostly is due to differences in effect sizes. Sample sizes are not adjusted accordingly across phenomena, but remain close to the typical sample size of 17.

Our first recommendation is thus to assess in advance how many participants would be needed to detect an effect. Note that we based our power estimations on whole meta-analyses, an analysis approach most suitable to make general statements about the status quo. 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 and for each as much detail pertaining procedure and measurements have been coded as possible (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; see for example Frank et al., 2016; ManyBabies Collaborative, 2017).

2. Carefully consider method choice. One way to increase power is the use of more sensitive measurements; as mentioned above, we do find striking differences between methods. When possible, it can thus be helpful to consider the paradigm being used, and possibly use a more sensitive way of measuring infants' capabilities. One reason that researchers do not choose the most robust methods might 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 and experience in how to interpret effect size estimates and use them for study planning (Mills-Smith, Spangler, Panneton, & Fritz, 2015).

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. Reports and discussions of effect sizes in experimental studies are rare, but despite long-standing 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).

A second impediment to meta-analyses in developmental science are current reporting standards, which make it difficult and at times even impossible to compute effect sizes from the published literature. For example, for within-participant measures it is necessary to report the correlation between conditions if two types of results are reported (most commonly outcomes of a treatment and control condition). However, this correlation, necessary to both compute effect sizes and their variance, is habitually not reported and has to be obtained via direct contact with the original 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 APA 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 necessary computations and opening the door for other types of cumulative analyses, for example in direct replications comparing raw results.

How to 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. 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 (Tsuji et al., 2014).

To support the improvement current practices, we propose to make meta-analyses available in the form of ready-to-use online tools, dynamic reports, and as raw data. These different levels allow researchers with varying interest and expertise interests to make the best use of the extant record on language development, including study planning by choosing robust methods and appropriate sample sizes. There are additional advantages for interpreting single results as well as for theory building that emerge from our collection of meta-analyses: On one hand, researchers can easily check whether their study result falls within the expected range of outcomes for their research question – indicating whether or not a potential moderator influenced the result. On the other hand, aggregating over many data points allows for the tracing of emerging abilities over time, quantifying their growth, and identifying possible trajectories and dependencies across phenomena (for a demonstration see Lewis et al., 2016). Finally, by making our data and source code open, we also invite contributions and can update our data, be it by adding new results, file-drawer studies, or new datasets. Our implementation of this proposal is freely online available at http://metalab.stanford.edu.

Cumulative evidence to decide whether skills are "absent" or not

Developmental research often relies on interpreting both significant and non-significant findings, particularly to establish a developmental time-line tracing when skills emerge. This approach is problematic for multiple reasons, as we mentioned in the introduction. 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 in fact impossible based on p values. Further, we want to caution researchers against interpreting the difference between significant and non-significant findings without statistically assessing it first (Gelman & Stern, 2006).

Concretely, we recommend the use of meta-analytic tools as demonstrated in this paper as well as in the work by Lewis et al. (2016). Aggregating over multiple studies allows not only for a more reliable estimate of an effect (because any single finding might either be a false positive or a false negative) but also makes it possible to trace developmental trajectories. A demonstration of such a procedure is given in the work of Tsuji & Cristia (2014) for native and non-native vowel discrimination. Their results match well with the standard assumption that infants begin to tune into their native language at around six months of age. For a contrasting example, see Bergmann & Cristia (2016), where the typically assumed developmental trajectory for word segmentation from native speech could not be confirmed, as across all included age groups infants seem to be able to detect words in the speech stream – the effect size of this skill is simply comparatively small and thus it is difficult to detect (see also Bergmann, Tsuji, & Cristia, 2017 for a more recent discussion of both meta-analyses).

Future directions

The present analyses can be expanded and improved in a number of ways. First, the present collection of meta-analyses does not represent an exhaustive survey of phenomena in language acquisition, let alone child development research. Particularly, topics typically investigated in younger children are over-represented. However, we sampled in an opportunistic, and thus to some degree random fashion, which lends some credibility to our approach. It would nonetheless be advisable to follow up on this report with a larger sample. To this end, we made all source materials along with extensive documentation available online.

Second, it would be important to further investigate the role of participant age in child development research. It is possible that developmental psychologists working with older age groups might focus on different issues or find that power and experimental design choices are less problematic; for instance, it may be easier to recruit larger samples via institutional testing in schools, and older children may be more reliable and consistent in their responses (Roberts & DelVecchio, 2000). We thus hope particularly to analyze more studies of older children to test this assumption.

Conclusion

We have demonstrated the use of standardized collections of meta-analyses for a diagnosis of (potential) issues in developmental research. Our results point to an overall lack of consideration of meta-analytic effect size in experiment planning, leading to habitually under-powered studies. In addition, method choice and participant age play an important role in the to be expected outcome; we here provide first estimates of the importance of either factor in experiment design. Assessing data quality, we find no evidence for questionable research practices and conclude that most phenomena considered here have evidential value. To ensure that developmental research is robust and that theories of child development are built on solid and reliable results, 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*, 1088–1101.

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *19*(6), 901–917.

Bergmann, C., Tsuji, S., & Cristia, A. (2017). Top-down versus bottom-up theories of phonological acquisition: A big data approach. *Submitted*.

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*. 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*(5), 365–376.

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*(4), 352–366.

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

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

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2016). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*.

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), 1–17.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124.

Jennions, M. D., & Møller, A. P. (2002). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, *269*(1486), 43–48.

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

Lewis, M. L., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., & Frank, M. C. (2016). A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis. *Preprint*. Retrieved from https://osf.io/htsjm/

ManyBabies Collaborative. (2017). Quantifying sources of variability in infancy research using the infant-directed speech preference. *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*(4), 416–432.

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*(6), 615–631.

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/

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*(1), 3.

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*(11), 1359–1366.

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191.

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/