Original Article

Addressing Publication Bias in Meta-Analysis

Empirical Findings From Community-Augmented Meta-Analyses of Infant Language Development

Sho Tsuji[1,2], Alejandrina Cristia[2], Michael C. Frank[3], and Christina Bergmann[4]

[1]International Research Center for Neurointelligence, Institutes for Advanced Studies, The University of Tokyo, Japan
[2]Ecole Normale Supérieure, Laboratoire de sciences cognitives et de psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[3]The Stanford Language and Cognition Lab, Department of Psychology, Stanford University, Stanford, CA, USA
[4]Language Development Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract: Meta-analyses are an indispensable research synthesis tool for characterizing bodies of literature and advancing theories. One important open question concerns the inclusion of unpublished data into meta-analyses. Finding such studies can be effortful, but their exclusion potentially leads to consequential biases like overestimation of a literature's mean effect. We address two questions about unpublished data using MetaLab, a collection of community-augmented meta-analyses focused on developmental psychology. First, we assess to what extent MetaLab datasets include gray literature, and by what search strategies they are unearthed. We find that an average of 11% of datapoints are from unpublished literature; standard search strategies like database searches, complemented with individualized approaches like including authors' own data, contribute the majority of this literature. Second, we analyze the effect of including versus excluding unpublished literature on estimates of effect size and publication bias, and find this decision does not affect outcomes. We discuss lessons learned and implications.

Keywords: meta-analysis, developmental psychology, effect sizes, gray literature

Meta-analyses are an indispensable research synthesis tool for characterizing bodies of literature and advancing theories. In typical meta-analyses, noisy measurements from multiple independent samples are normalized onto a single scale (typically a measure of effect size) and combined statistically to produce a more accurate measurement. Effects for meta-analysis can come from the published literature, unpublished data, or even the author's own work, but different strategies for identifying datapoints for inclusion can have major consequences for the interpretation of the meta-analytic estimate. In particular, the exclusion of unpublished work can lead to a bias for positive findings and hence compromise validity. Thus, it is important to assess the utility – and impact – of strategies for including unpublished data. In the present article, we describe our successes and failures with gathering unpublished data for meta-analyses within developmental psychology, and assess how the addition of these datapoints changes the conclusions from our sample of meta-analyses.

Community-Augmented Meta-Analyses and MetaLab

Community-augmented meta-analyses (CAMAs, Tsuji, Bergmann, & Cristia, 2014) are a tool for countering some problems faced by traditional meta-analyses. In the original proposal, CAMAs were imagined as open-access, online meta-analyses: living documents that can be openly accessed, updated, and augmented (Tsuji et al., 2014). Their dynamic nature avoids a key problem of traditional meta-analyses, which are crystallized at the time of publication and quickly become outdated. Additionally, CAMAs were set up to allow the addition of unpublished datapoints. Although we initially aimed for authors and others to add studies to extant meta-analyses, we now favor a system where a single curator is responsible for updating a given meta-analysis. This preserves the original goal of having up-to-date meta-analyses, and further ensures internal consistency in all meta-analyses. This change in the concept of curation (from crowd-sourcing to centralized), however,

does not affect the topics that are broached in this paper, and thus will not be discussed further.

MetaLab is a database and browsable web interface that instantiates the CAMA idea (http://metalab.stanford.edu/; Bergmann et al., 2018). The database's focus is Developmental Psychology, and the goal is to eventually cover all subfields on which there are experimental results bearing on infant and child cognition. At present, MetaLab hosts 20 meta-analyses (containing a total of 1,686 effect sizes), covering diverse topics ranging from sensitivity to vowel contrasts (e.g., the sound difference between "ship" and "sheep"; Tsuji & Cristia, 2014) to children's preference for prosocial over anti-social agents (Margoni & Surian, 2018). Most meta-analyses, however, bear on language development, and focus on children aged 5 years or younger.

In the present paper, we analyze 12 meta-analyses in MetaLab for which efforts like search strategy and contact with authors were well documented and accessible to us (containing a total of 1,232 effect sizes; Bergmann & Cristia, 2016; Black & Bergmann, 2017; Carbajal, 2018; Cristia, 2018; Fort et al., 2018; Rabagliati, Ferguson, & Lew-Williams, 2019; Tsuji & Cristia, 2014; Tsui, Byers-Heinlein, & Fennell, 2019; Von Holzen & Bergmann, 2018). Some of these meta-analyses were co-authored by authors of the present article. We discuss below to what extent our results may generalize to other meta-analyses and fields of psychology.

Unpublished Literature in Meta-Analyses

Since meta-analyses largely build on publicly accessible literature, they face some of the same challenges as primary literature in the context of the replication crisis (Lakens et al., 2017). One key issue concerns the inclusion of unpublished data, that is, results that do not appear in the published literature (and hence may not be indexed by all libraries and academic search engines), but are either reported in theses, dissertations, conference abstracts, white papers or internal reports, or not reported publicly at all (i.e., studies that are "file-drawered").

Attempting to access unpublished data is difficult and time-consuming. To begin with, reports on these data, if they exist, tend to not be indexed as carefully as published data and thus are harder to discover. For instance, a search on PubMed would not reveal theses or dissertations, whereas Google Scholar does index some (but not all) thesis archives. Even if a meta-analyst uses Google's Scholar engine, conference abstracts and proceedings in many fields are not indexed, and thus need to be searched manually. In some cases, for instance when conferences in a field favor very short abstracts, one may discover the existence of a study, but be unable to integrate it because there is insufficient information reported. In this case, as well as in the case of studies for which reports do not exist, author contact is the only way to secure the information needed to integrate a study into a quantitative analysis.

One may try to write to all authors who have published on the topic, and ask for data in their file drawers. This is likely a biased approach, however, since authors of file-drawer studies that have never published on the topic cannot be accessed by this strategy. Those authors might, however, be the ones that have collected and failed to successfully published data that go against the main direction of findings for the field. To work against such biased collection and access also data collected by others, one can publicly post a call for data, for example, via field-specific mailing lists. Thus, meta-analysts who intend to include gray literature can be led to make a significant investment in time to be able to discover and integrate such results. To our knowledge, there is no previous research documenting the effectiveness of these modes of gray literature integration for psychological research. Therefore, in Part one below, we have undertaken to document the efficacy of these diverse methods, (i) database searches, (ii) citation searches, (iii) mailing list calls, (iv) cases where authors' work was known (v) inclusion of own data. Relatedly, we also document the success rate in gathering data based on emailing authors with a request for information.

Although discovering and integrating unpublished data is costly, it is often part of standard meta-analytic practice recommendations (e.g., White, 1994), in the hope that it will reduce publication bias. Indeed, published literature is widely assumed not to constitute an unbiased sample of the data, in turn yielding an overestimation of effect sizes in meta-analyses that only include published literature (e.g., Guyatt et al., 2011). Ferguson and Heene (2012) note that at least 25% – and possibly as many as 80% – of meta-analyses in psychology suffer from significant bias. A vast body of evidence confirms that this is a concern for psychological science in particular. For instance, Bakker, van Dijk, and Wicherts (2012) show convincingly that researchers in psychology typically use small sample sizes (with an inordinate proportion of statistically significant results), rather than larger sample sizes (whose higher precision reduces the likelihood of false positives and negatives). The problem is so widespread that item 15 of the PRISMA checklist specifically asks meta-analysts to "Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias...)" (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009) and a systematic review of systematic reviews on the effects of all sources of bias

identifies the inclusion of gray literature among its key recommendations (Tricco et al., 2008).

Most meta-analyses, and therefore most recommendations for meta-analyses, are based on the medical intervention literature, however. Studies of publication biases from this field may or may not generalize to psychological research. One previous study investigated bias and unpublished data inclusion for 91 meta-analyses published in psychological journals (Ferguson & Brannick, 2012). Surprisingly, they concluded that meta-analyses including unpublished data were more, rather than less, biased than studies based purely on published data. These authors recognize the validity of the gray literature inclusion approach for medical meta-analyses, where registries allow for unbiased discovery of studies, and mandatory preregistration of studies further precludes analyses that favor specific results (Huić, Marušić, & Marušić, 2011). Since neither of these factors exist for psychology, it may be unwise for psychological meta-analyses to include gray literature because (1) the effort will be too large for the number of effect sizes that can be included ultimately (with a median of fewer than 5% of effect sizes stemming from unpublished data; Ferguson & Brannick, 2012); and (2) unpublished data will be biased because they are discovered mainly via a biased network: the meta-analysis' authors and close colleagues, and prominent authors in the field, all of whom may contribute data that favor a given outcome. In view of these contrasting results between the psychological literature and the body of meta-analytic best practices research, we revisit the question of what the effects of adding unpublished data are based on our CAMAs.

In Part two, we follow previous literature (e.g., Tricco et al., 2008) and report: (i) effect size estimates for samples with and without unpublished literature; (ii) bias estimates with and without unpublished literature; and (iii) potential correlates of study quality. We note that study quality is much harder to measure objectively in basic psychological research than in interventions. In interventions, randomized control trials with a double-blind procedure are undoubtedly better quality evidence for causal links than correlational research. Such hierarchy can be harder to establish for some types of laboratory experiments, where procedures like experimenter blinding or randomization exist, but might be implemented much less systematically and consistently than in intervention studies. However, we can at least inspect some general features that may correlate with data quality, for instance a study's sample size. Some previous work suggests that unpublished data are lower quality by being based on smaller samples (e.g., Tricco et al., 2008).

Finally, we dedicate a third part to in-depth case studies and summaries of lessons learned.

Methods

All data reported in the Results section will be based on a subset of meta-analyses openly available on MetaLab. We include those meta-analyses that are based on a systematic literature search, made efforts to include unpublished data, and documented their data gathering efforts systematically. We define as unpublished anything that is not in a peer-reviewed journal, including work that has appeared only in theses, proceedings, and books or book chapters.

Cohen's d is a standardized effect size based on sample means and their variance. We rely here on d values computed in the MetaLab pipeline, which uses standard formulae to convert the measurements reported in papers to d values (details are reported elsewhere; Bergmann et al., 2018, see also http://metalab.stanford.edu/).

Data and analyses scripts are shared on our Open Science Framework project site (https://osf.io/g6abn/) and on PsychArchives (data: https://www.psycharchives.org/handle/20.500.12034/2185, code: https://www.psycharchives.org/handle/20.500.12034/2186). Analyses have been conducted with the tidyverse (Wickham, 2017) and the metafor (Viechtbauer, 2010) packages in R (R Core Team, 2019).

Results

Of the 20 meta-analyses included in MetaLab at time of writing, 14 meta-analyses (70%) include unpublished data. This proportion is comparable to previous reports, where 63% of recent meta-analyses in Psychology made efforts to include gray literature (Ferguson & Brannick, 2012). Of those meta-analyses that did not restrict their search to published data, 12 fit our additional criteria for inclusion in the present analysis, namely being based on a systematic literature search, and systematically documenting data gathering efforts and/or making those efforts accessible to us. Concretely, the meta-analyses included in our final sample needed to have made available their search procedure in a document and/or provided it to us for the purpose of the present studies. A literature search was deemed systematic if it included and documented a keyword or seed search and details on the databases searched and search dates. Authors further needed to have documented a number of records found and included, their inclusion and exclusion criteria, as well as provide an exhaustive list of other sources consulted to gather information and data. For the purpose of the present study, we aggregate two pairs of meta-analyses into single meta-analyses, since the systematic literature review in both cases had originally been conducted on the pair, and the datasets were only

Table 1. Overview of meta-analyses and number of published and unpublished effect sizes

				N unpublished effect sizes					
Citation	Meta-analysis	*N* papers	*N* effect sizes	Overall	Database search	Citation search	Author known	Own	*N* data obtained through author email
Carbajal (2018)	Familiar word recognition	15	33	12	2	4	4	2	4
Bergmann and Cristia (2016)	Natural word segmentation	73	315	25	23	0	0	2	12
Von Holzen and Bergmann (2018)	Mispronunciation sensitivity	32	251	32	0	0	33	0	14
Black and Bergmann (2017)	StatSeg	26	91	10	10	NA	0	0	2
Tsuji and Cristia (2014)	Vowel discrimination	38	194	11	4	NA	6	1	6
Fort et al. (2018)	Sound symbolism	11	44	20	2	NA	10	8	6
Rabagliati et al. 2019	Abstract rule learning	20	94	4	4	NA	0	NA	3
Cristia (2018)	Phonotactic learning	15	47	11	0	0	1	10	2
Cristia (2018)	Statistical sound category learning	11	20	9	0	0	5	2	2
Tsui et al. (2019)	Switch task	47	143	11	2	0	0	9	3
	Mean	28.8	123.2	14.4	4.7	0.4	5.9	3.4	
	% of unpublished effect sizes				37.7	3.3	31.8	27.5	

later thematically separated in MetaLab.[1] The resulting 10 meta-analyses ranged in size from 11 to 73 papers (*Mdn* = 23), with 20–315 datapoints included (*Mdn* = 92). Table 1 gives an overview of the meta-analyses, including citations, descriptive statistics on the number of effect sizes by publication status, and how this unpublished literature was found.

Evaluation of Data-Gathering Efforts

Overall, of the total of 1,232 effect sizes contained in our MetaLab subset, 144 effect sizes, or 11.7% of data, were based on unpublished literature. Similar to previous reports (Ferguson & Brannick, 2012), the distribution of unpublished study percentages shows a positive skew, with most meta-analyses in our sample having a low percentage of unpublished studies – around 10% or less (Figure 1).

Database Searches

All database searches followed standard meta-analytic practice, wherein a set of pre-determined keywords were entered into a search engine, and titles of hits and the abstracts of potentially relevant papers were scanned to arrive at the sample of papers eligible for inclusion in the respective meta-analysis. In all meta-analyses included in our study, Google Scholar was the search engine of choice, which performs equivalent to a combined search of multiple databases (Gehanno, Rollin, & Darmoni, 2013), although exact replicability of Google Scholar searches might be compromised since it saves users' history.

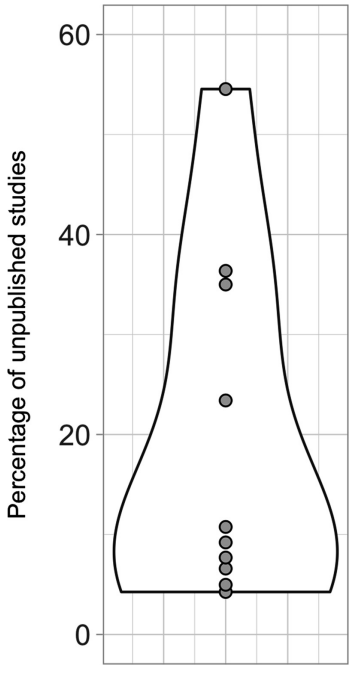

Figure 1. Percentage of unpublished studies included per meta-analysis.

Crucially for us, Scholar includes unpublished work (i.e., pre-prints, conference proceedings, or unpublished manuscripts) in its search results as long as they are available online and indexed. We found that an average of 4.7 datapoints or 37.3% (*SD* = 42.9) of unpublished data was found by Google Scholar searches (Figure 2).

[1] "Word segmentation" and "Function word segmentation" are aggregated into "Natural word segmentation"; "Native vowel discrimination" and "Non-native vowel discrimination" are aggregated into "Vowel discrimination".

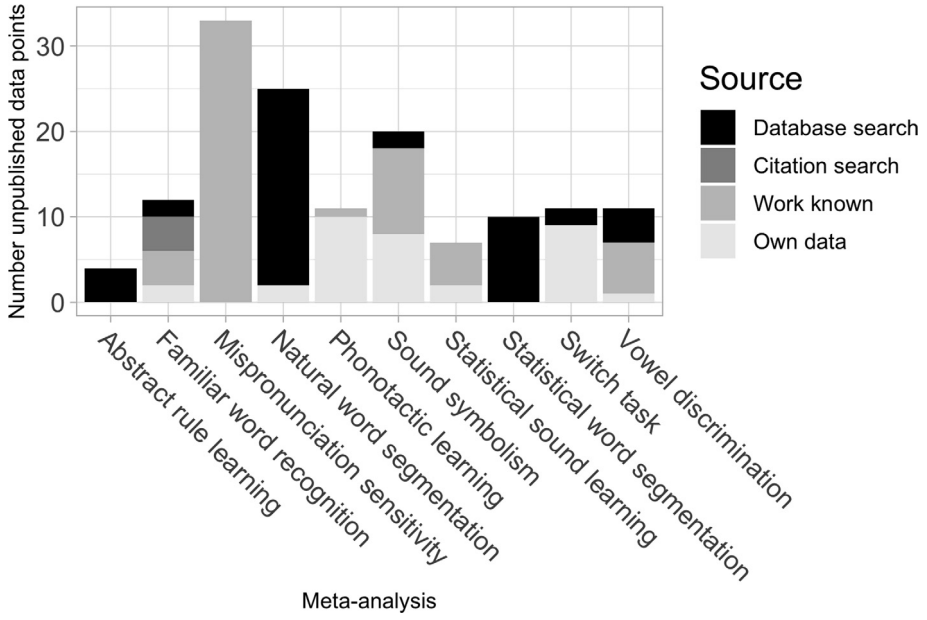

Figure 2. Number of unpublished datapoints obtained through different sources of recovering gray literature, by meta-analysis.

Citation Searches

Some unpublished studies might not be available online and thus not detectable by search engines. They might, however, be discoverable by searching the reference lists of available studies. The percentage of unpublished datapoints gathered based on citation searches was an average of 0.4 datapoints or 3.3% (*SD* = 10.5).

Mailing List Requests

In order to reach a relevant audience to recover potential gray literature, authors of six of the included meta-analysts requested contributions via professional email lists. Strikingly, these attempts did not lead to a single reply with information that could be added to the meta-analyses.

Author's Work Known

In addition to the more formal routes described above, a meta-analyst can get to know an author's eligible work at a conference, or via informal communication with experts in the field. Our estimate of datapoints added via this route is an average of 5.9 datapoints or 31.8% (*SD* = 35.9). Since a meta-analyst is often an expert in the topic of their meta-analysis, this can be a very fruitful route – one of our special cases below will illustrate how helpful this strategy can be for the data gathering process.

Own Data

A meta-analyst might also contribute their own unpublished data to their meta-analysis. In MetaLab, unpublished data from meta-analysts' own research accounted for an average of 3.4 datapoints or 27.5% (*SD* = 33.7) of total unpublished datapoints. Previous reports have documented a difference in own published and unpublished contributions, with an average of 5.89% of total published datapoints, and 12.94% of total unpublished datapoints, being based on meta-analysts' own data (Ferguson & Brannick, 2012). If we also assess published datapoints with this metric, we find that an average of 9.6% of published datapoints are based on own data in MetaLab. Although this ratio suggests more addition of unpublished own datapoints, note, that there are much more total published than unpublished datapoints in the MetaLab datasets. If we look at the absolute number of datapoints added, meta-analysts added more published (an average of 6.1) than unpublished (an average of 3.4) own datapoints to their datasets. Given that the absolute number of published versus unpublished datapoints in the previous literature (Ferguson & Brannick, 2012) is comparable or lower than what is found the present study, we can conclude that MetaLab contains a relatively high proportion of authors' own unpublished data.

Emails to Authors

Meta-analysts can chose to contact authors of papers eligible for inclusion in their meta-analysis with request for additional information from eligible literature (whether published or unpublished), and whether they are aware of any gray literature. It was impossible to recover the number of effect sizes added based on these requests, since meta-analysis authors did not consistently document how many datapoints of a given study were affected by their request (e.g., only one experiment of a study or all experiments could have been affected), and whether the information

gathered was necessary to compute published or unpublished effect sizes. We therefore instead counted the number and ratio of authors contacted, and how many of these contacts were responsive and lead to data that could be added to the meta-analysis. An average of 12.7 authors was contacted. Out of all authors contacted, an average of 9.8 or 85.1% (*SD* = 19.5) were responsive and 5.4 or 49.6% (*SD* = 28.1) provided data that could be added to the meta-analysis.

Community Contributions

Although the original CAMA idea entailed that, ultimately, the research community would take over the curation of datasets and infrastructure in a bottom-up fashion, this model has not proven feasible. Two issues were a general lack of community contributions and difficulty in curating entries that contained errors. Instead, MetaLab now consists of a governing board that, aided by external funding,[2] maintains and expands the general infrastructure. Dedicated curators for each dataset are then in charge of updating individual meta-analyses.

Meta-Analyses With Published and Unpublished Results

The second part of our analysis evaluated the effect of including gray literature on publication bias. We evaluate the effect of including or excluding unpublished literature on effect size and bias estimates in our samples. We also assess potential correlates of study quality.

Effect Size Estimates by Publication Status

The mean Cohen's *d* effect size across meta-analyses is d = 0.22 (range: −0.34 to 0.77). The mean is d = 0.24 (range: −0.17 to 0.66) for published, and d = 0.15 (range: −0.34 to 0.77 for unpublished studies, consistent with publication bias (greater effects for journal-published studies) as well as a difference in data quality (lower for not journal-published studies). This analysis ignores factors that are known to vary across studies, however.

Therefore, in order to more specifically assess the effect of inclusion of unpublished datapoints into a meta-analysis, we constructed meta-analytic regression models for the full dataset as well as datasets including only the published or unpublished datapoints (see, e.g., Tricco et al., 2008). The model for each meta-analysis included infant age as a predictor, since this factor has been consistently found to explain variance in effect sizes (Bergmann et al., 2018). While testing method is another such factor, we refrained from including it in our analyses, since data subsets differ in the number of testing methods included, with some subsets being comprised of data stemming from only one testing method. Our random effects structure allowed shared variance for datapoints stemming from the same paper, and accounted for the dependence between datapoints stemming from the same infant participants contributing multiple effect sizes (see Konstantopoulos, 2011; R model = `rma.mv(d, d_var, mods = ~ age, random = ~ 1| paper/same_infant/data_point)`). We had to exclude one data subset for which the regression model did not converge. Note that meta-analytic regression estimates become imprecise with small datasets, and we could have addressed this issue by excluding those data subsets with small amounts of datapoints. However, since any such cut-off would be arbitrary, we opted for including regression analyses for all datasets that did converge.

Figure 3 shows the resulting effect size estimates and associated confidence intervals. There is no clear pattern in terms of higher meta-analytic effect size estimates for published datasets, consistent with previous reports (e.g., Chow & Ekholm, 2018; Guyatt et al., 2011), and confidence intervals for the respective sets mostly overlap. Thus, when known factors structuring variance (age, method, meta-analysis) are accounted for, there does not seem to be a clear pattern as to the direction in which the inclusion of gray literature affects meta-analytic conclusions. Note, however, that effect sizes do change based on the inclusion of gray literature for the majority of datasets, and that including or excluding these studies would likely affect the overall conclusions of any given meta-analysis. Given the overall small sample sizes, it is impossible to estimate to what extent the fact that we include specifically gray literature – as opposed adding literature in general – affects these effect size estimates.

Bias Estimates With and Without Unpublished Literature

In order to evaluate the impact of inclusion or exclusion of unpublished literature on bias estimates, we assessed each individual meta-analysis by means of funnel plot asymmetry, a classical diagnostic for identifying potential publication bias (Egger, Smith, Schneider, & Minder, 1997). We included as moderators infant age, a factor that explains variance in most meta-analyses in MetaLab. The distribution of test statistics for Egger's test for funnel plot asymmetry did not differ whether we assessed datasets under exclusion of unpublished studies (z_{min} = −1.89, z_{mean} = 3.33, z_{max} = 13.41), or when gray literature was included (z_{min} = −1.90, z_{mean} = 3.20, z_{max} = 12.44) (see Figure 4). In both subsets, the same 4 out of 10 datasets showed significant funnel plot asymmetry, suggesting that adding gray

[2] https://www.bitss.org/projects/metalab-paving-the-way-for-easy-to-use-dynamic-crowdsourced-meta-analyses/

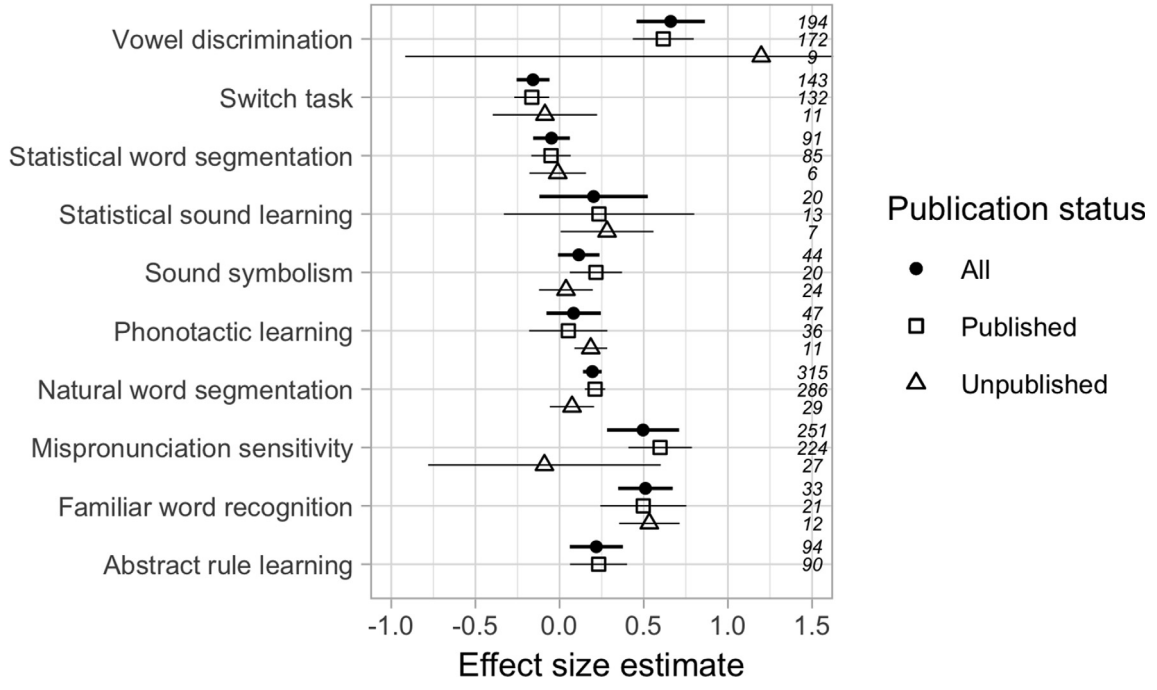

Figure 3. Mean meta-analytic effect size estimates and associated confidence intervals. Values are based on meta-analytic regression models for each dataset. Shapes represent different subsets per meta-analysis based on their publication status, lines indicate the 95% CI. Numbers in italics on the right side indicate the number of effect sizes going into each regression model.

literature did not improve this indicator of publication bias. A more quantitative evaluation of this difference proved difficult. For instance, a non-parametric bootstrapping approach is unreliable for meta-analyses with fewer than about 50–100 studies, which is the case for the majority of the meta-analyses included in the present analysis. Although funnel plot asymmetry is a classic diagnostic for assessing publication bias, it tends to have low power and can fail to detect significant publication bias (e.g., Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006; Macaskill, Walter, & Irwig, 2001).

In order to evaluate the effect of studies' publication status more directly, we ran a random-effects meta-analytic regression for each meta-analysis, using publication status (published, unpublished) as a moderator. We again included infant age and testing method as additional moderators. We assessed whether publication status had a significant effect in each of these meta-analyses, and found that this was the case in two cases. In both cases, gray literature had a negative impact on effect sizes (see Table 2). Thus, even though including gray literature did not cause a significant change in the results of Egger's test for publication bias, it did explain a significant proportion of variance in several datasets. Further, gray literature inclusion may also be relevant in other cases; it is plausible that we failed to detect such an effect due to the small samples included in our meta-analyses leading to low statistical power.

Correlates of Study Quality

There are no uniformly agreed-upon markers of study quality in experimental psychology research that can easily be assessed. Since unpublished effect sizes are often based on smaller samples (Tricco et al., 2008), we assessed whether sample size could serve as a proxy of quality. However, a descriptive assessment of sample size revealed no difference in the sample size for published (M = 21.7, SD = 9.9) and unpublished studies (M = 22.5, SD = 10.3) in our sample of meta-analyses.

In addition to sample sizes, we attempted to assess one other potential indicator of study quality, namely, internal correlations in within-participant designs. Since weighted meta-analytic regression requires an estimate of these correlations, some meta-analysts have gathered this measure. A higher internal correlation might suggest less noise in the measure, thus potentially indicating higher study quality. We first checked whether the degree of internal correlation correlated positively with child age (since measure precision tends to improve as children age). However, our assessment showed that internal correlation and child age were negatively correlated significantly [$r = -0.28$, $t(417) = -5.88$, $p < .001$]. Since this result contradicted our initial assumption, we did not pursue this possibility further.

Overall, therefore, we were not able to show any relationship between potential measures of study quality and

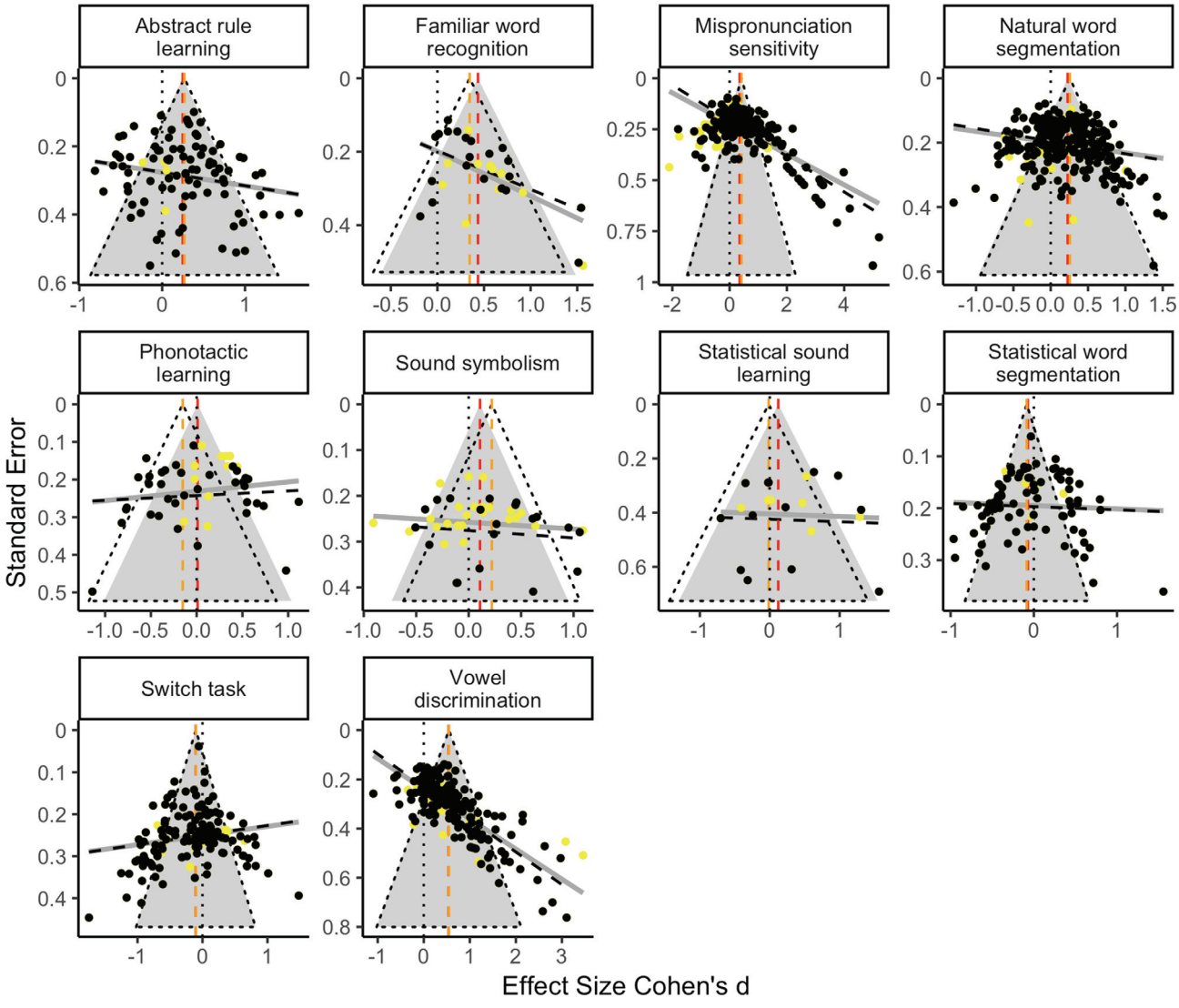

Figure 4. Funnel plots by dataset. Effect size estimates for published data are represented by black points, and estimates for unpublished data by yellow points in the color version of this figure available with the online version of this article. The mean effect size for the full dataset is shown as a red dashed line (invisible when the means for full and published data sets overlap), and the gray shaded funnel corresponds to a 95% CI around this mean. The mean effect size for the subset of published studies is shown as an orange dashed line, and the transparent dashed funnel corresponds to a 95% CI around this mean. The gray dashed line shows an effect size of zero. In the absence of bias, we should expect all points to fall symmetrically inside the funnel.

publication status, likely due in part to the lack of objective criteria for study quality in experimental psychology.

Case Studies

Case Study 1: Expanding the Pool of Meta-Analyses

In addition to gathering unpublished data and missing datapoints for extant CAMAs, it was also sought to expand the pool of CAMAs available on MetaLab. For this purpose, a call for contributions to MetaLab was issued, with a $1,000 cash prize for the top three most extensive meta-analyses submitted. To advertise this challenge, announcements were sent to professional mailing lists, the literature was searched for extant meta-analyses fitting the scope and their authors were contacted. Information on ongoing meta-analyses efforts were informally gathered and distributed. These efforts resulted in six eligible submissions for the challenge, four of which are already integrated in MetaLab. Considering that these will finally constitute 27.2% (6/22) of meta-analyses on MetaLab, this strategy substantially expanded the database at relatively low cost compared with the cost of performing new meta-analyses.

Table 2. Meta-analytic regression coefficients for the effect of publication status by dataset

	Effect sizes	β	*SE*	*z*	*p*	ci.lb	ci.ub
Abstract rule learning	94	−0.051	0.178	−0.284	.776	−0.399	0.298
Familiar word recognition	33	0.055	0.088	0.626	.532	−0.118	0.228
Mispronunciation sensitivity	251	−0.476	0.144	−3.297	.001*	−0.758	−0.193
Natural word segmentation	315	−0.071	0.044	−1.638	.101	−0.157	0.014
Phonotactic learning	47	0.068	0.09	0.762	.446	−0.107	0.244
Sound symbolism	44	−0.127	0.062	−2.032	.042*	−0.249	−0.005
Statistical sound learning	20	−0.01	0.171	−0.056	.955	−0.345	0.326
Statistical word segmentation	91	0.02	0.111	0.185	.854	−0.197	0.237
Switch task	143	0.036	0.089	0.402	.688	−0.138	0.209
Vowel discrimination	181	0.177	0.209	0.848	.397	−0.232	0.586

Note. *SE* = standard error; ci.lb = lower boundary of confidence interval; ci.ub = upper boundary of confidence interval. Asterisk indicates statistical significance.

Case Study 2: Gathering Gray Literature Through In-Person Author Contact

One of the MetaLab datasets on sound symbolism in infancy (Fort et al., 2018) assesses the development of the bouba-kiki effect, whereby humans associate pseudowords like "bouba" with round objects, and pseudowords like "kiki" with spiky objects. Experiments examining this effect have had mixed results in infant populations. Anecdotally, several researchers have failed to find a consistent bouba-kiki effect, but have faced problems publishing these null results. Encountering others researchers with similar results through conference presentations, they encouraged each other to share their unpublished data and decided to conduct a meta-analysis on the phenomenon with both published and unpublished results. This meta-analysis reveals that there is overall evidence for a bouba-kiki effect in infants, however, it is smaller than suggested by the published literature alone. This small effect size, combined with habitually small sample sizes in infants' studies, likely explains the divergence in findings between attempts to elicit the bouba-kiki effect. In the case of this meta-analysis, presenting null results at a relevant conference, and thus making others aware of their existence, proved a highly effective way to assemble gray literature.

Discussion

Publication bias is considered a key problem of the meta-analytic literature, and the exclusion of gray literature from meta-analyses is a potential cause. Since the difficulty of accessing such gray literature is a reason for the lack of unpublished studies in meta-analyses, we assessed the amount of gray literature gained based on various strategies in datasets assembled in the open-access database MetaLab (Bergmann et al., 2018). In the following, we will discuss these strategies from the viewpoint of lessons learned and recommendations for future meta-analysts. We further assessed the impact of including such gray literature on publication bias. These analyses show that our efforts had only a moderate impact on publication bias in our datasets, a result we will discuss in light of previous literature and the nature of gray literature gathered in our datasets.

Lessons Learned From Efforts Gathering Missing Data

If an article does not report all data necessary for estimating effect sizes, contacting the original authors of the article is the only way to possibly obtain these missing data. Although this is an effortful endeavor, our analysis of data gathering efforts shows that it is a successful strategy to gather missing information. Authors contacted individually by email were highly responsive and sent data useful for computing effect sizes in almost half of the cases. Of course, no reply was forthcoming in the other half of cases, and the fact that this outcome should still be considered a success illustrates the difficulty of data gathering during meta-analyses, especially considering that our mailing list calls failed to point us to any missing data. While we do not have comparative data for other approaches, the successful author contacts by meta-analysts in MetaLab are based on highly individualized emails to the respective first and/or last authors of an article. Along with outlining the general aim of the meta-analysis, meta-analysis authors would mention both the authors' and their article's name, and explain in detail the nature of data needed from them. Habitually, we would send one follow-up email in case we did not get a reply. In addition to these efforts, it has been shown that adding data-sharing agreements to requests for primary data can improve responsiveness (Polanin & Terzian, 2019).

Lessons Learned From Efforts Gathering Gray Literature

On average, 11.4% of effect sizes were based on unpublished studies in the present dataset. Standard tools like database searches, which arguably take less time and effort than more personalized efforts like knowing of an author's work, can already contribute an important amount of gray literature. Although it is difficult to estimate the exact time effort required, some gray literature would be included in the search automatically depending on the engine chosen, and thus the search itself would not require more time investment. Another standard tool, citation searches, only lead to the discovery of unpublished studies in one database. Although such searches are an added effort, a meta-analyst would be recommended to carefully read the included literature, which is not too many steps away from the citation search itself.

Knowing the author of a study (whether in person or not), proved to be a fruitful strategy to gather data. Although this strategy might be more dependent on an individual meta-analysts' network and the availability of unpublished studies in conferences and other places enabling personal contact, our analyses and case study on sound symbolism illustrate that this is a promising way to include unpublished studies, and should be on the meta-analyst's mind as a possible way to gather data. Ideally, though, we think it would be desirable that there were better indexing of unpublished literature (by authors and conference organizers uploading their unpublished work to searchable archives). A more reliable index of gray literature would reduce the individual's dependence on high effort strategies and enable discovery via the standard database search strategies, which at the same time increases transparency. Finally, even to a greater extent than the previous literature (Ferguson & Brannick, 2012), the meta-analyses included here contained a relatively high proportion of effect sizes stemming from the meta-analysts' own data. A meta-analysts' own unpublished data, or such data in their network, might induce bias, since owning such data might serve as a motivator to conduct an MA, or own data might be more likely included in an MA before peer review. Thus, such potential bias might counteract the otherwise bias-reducing effect of adding unpublished literature to an MA.

The Effect of Adding Unpublished Literature on Effect Sizes and Publication Bias

Previous literature has reported higher effect sizes among published effect sizes, potentially leading to an overestimation of effect sizes if unpublished literature is not included. In contrast, other authors have warned against the inclusion of gray literature because it would increase bias (Ferguson & Brannick, 2012). We addressed this question with two types of analyses. Our funnel plot analyses suggested that these problems only played a minor role in our data, showing no differences in publication bias by publication status. Our second analysis for addressing this issue, a meta-analytic regression with publication status as a moderator, showed that, for two datasets, adding unpublished literature had a significant or marginally significant impact on effect sizes.

Even the lack of statistical bias in one of our analyses does not necessarily mean that the unpublished literature we include is not biased, especially considering the relatively low power of our sample. As mentioned in the Introduction, detectable gray literature might itself be biased. Indeed, recent large-scale analyses of effect sizes in published versus unpublished studies included in meta-analyses indicated systematically larger effect sizes for the published literature (Polanin, Tanner-Smith, & Hennessy, 2016), and the fact that we do not find such a difference might indicate that our sample of unpublished studies is upwards biased. Similarly, study preregistration has been shown to be strongly associated with more null results (Kaplan & Irvin, 2015), suggesting that, without preregistration, the tendency to publish null findings is lower. Overall, our results thus suggest that unpublished studies should be added to a meta-analysis with great care and transparency, allowing the reader to gain insight into the effect these datapoints have on the overall estimates.

Adding gray literature might not be equally illuminating, necessary, or damaging in every field of psychology. With regard to the advantages and disadvantages of including gray literature in the case of infant literature, we suggest that such literature will improve the overall quality of a database, for at least two reasons. First, the infant literature is tremendously underpowered and benefits from a larger body of studies to better estimate true effect sizes. Second, in order to conduct infant experiments, a researcher habitually needs to undergo training and make use of a specialized laboratory, making it unlikely that unpublished data are especially prone to being badly designed or executed. Third, if gray literature is added the way we suggest in the context of CAMAs such that each study is coded based on publication status, meta-analysts and database users can decide for themselves whether or not to include unpublished datapoints in their assessments.

A relatively large amount of unpublished data in the present dataset was based on potentially biased data, namely the meta-analyst's own data or data based on direct contact with study authors. Potentially, reducing this bias in

the future might impact the difference between effect sizes and measures of publication bias between published and unpublished studies in the present dataset.

Limitations and Opportunities

Meta-analyses in MetaLab are not a random or representative selection of meta-analyses in psychology. Some of their characteristics are atypical of meta-analyses in Psychology; for instance, the number of studies included (*Mdn* = 23) is larger than the median of 12 reported in larger samples (Van Erp, Verhagen, Grasman, & Wagenmakers, 2017), and include a larger number of effects per study.

In fact, two of the included meta-analyses themselves have not yet been published (Carbajal, 2018; Tsui et al., 2019), and two more have been peer-reviewed for proceedings papers (Black & Bergmann, 2017; Von Holzen & Bergmann, 2018), and thus they have not, or to a moderate degree, been affected by the review process. The meta-analysis authors were often students who were doing a systematic review of a literature to which they were planning to contribute, and thus may not have had as much of a vested interest in supporting one or another theory as more established researchers might do. On the other hand, a relatively high proportion of unpublished data stemmed from meta-analysis authors, which might indicate a comparatively high interest in supporting a specific theory. Finally, most of them come from a cluster of researchers (including the authors of this paper) who strived to follow best practices guidelines such as following the PRISMA statement (Moher et al., 2009).

While these characteristics might limit comparability with other attempts, the transparent data gathering process by relatively unbiased meta-analysts might enable us to assume that the biases found in the datasets can be attributed to the literature itself more than to the meta-analytic process. Then, considering that all meta-analysts of datasets included in the present sets attempted to gather gray literature, but publication bias was still prevalent, the MetaLab subset re-emphasizes a broader problem of the field, namely the lack of publicly available indexing of gray literature.

Another characteristic of MetaLab is its basis on the CAMA approach, which from the outset was meant to function as a natural home for file-drawer studies in addition to published studies. Opening the file-drawer in this way can also help us to estimate how many studies are filtered out by the peer review process in the future. Metalab's growth to now 20 datasets indicates its success. The website's visibility has led to numerous conference presentations on the included meta-analyses as well as invitations to provide tutorials, which in turn have inspired others to start their own meta-analyses and become curators, or else to render more visible extant meta-analyses.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554. https://doi.org/10.1177/1745691612459060

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

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*, 1996–2009. https://doi.org/10.1111/cdev.13079

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 124–129). Austin, TX: Cognitive Science Society.

Carbajal, M. J. (2018). *Separation and acquisition of two languages in early childhood: A multidisciplinary approach*. (Unpublished doctoral dissertation). Ecole Normale Supérieure, Paris, France.

Chow, J. C., & Ekholm, E. (2018). Do published studies yield larger effect sizes than unpublished studies in education and special education? A meta-review. *Educational Psychology Review, 30*, 727–744. https://doi.org/10.1007/s10648-018-9437-7

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition, 180*, 312–327. https://doi.org/10.1016/j.cognition.2017.09.016

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*, 120–128. https://doi.org/10.1037/a0024445

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555–561. https://doi.org/10.1177/1745691612459059

Fort, M., Lammertink, I., Peperkamp, S., Guevara-Rukoz, A., Fikkert, P., & Tsuji, S. (2018). SymBouki: A meta-analysis on the emergence of sound symbolism in early language acquisition. *Developmental Science, 21*, e12659. https://doi.org/10.1111/desc.12659

Gehanno, J. F., Rollin, L., & Darmoni, S. (2013). Is the coverage of Google Scholar enough to be used alone for systematic reviews. *BMC Medical Informatics and Decision Making, 13*, 7. https://doi.org/10.1186/1472-6947-13-7

Guyatt, G. H., Oxman, A. D., Montori, V., Vist, G., Kunz, R., Brozek, J., … Williams, J. W. Jr. (2011). GRADE guidelines: 5. Rating the quality of evidence-publication bias. *Journal of Clinical Epidemiology, 64*, 1277–1282. https://doi.org/10.1016/j.jclinepi.2011.01.011

Huić, M., Marušić, M., & Marušić, A. (2011). Completeness and changes in registered data and reporting bias of randomized

controlled trials in ICMJE journals after trial registration policy. *PLoS One 6*, e25258. https://doi.org/10.1371/journal.pone.0025258

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One, 10*, e0132382. https://doi.org/10.1371/journal.pone.0132382

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

Lakens, D., van Assen, M. A. L. M., Anvari, F., Grange, J. A., Gerger, H., Hasselman, F., … Zhou, S. (2017). Examining the reproducibility of meta-analyses in psychology: A preliminary report. *BITSS Preprint*. https://doi.org/10.31222/osf.io/xfbjf

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal, 333*, 597–600. https://doi.org/10.1136/bmj.333.7568.597

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20*, 641–654. https://doi.org/10.1002/sim.698

Margoni, F., & Surian, L. (2018). Infants' evaluation of prosocial and antisocial agents: A meta-analysis. *Developmental Psychology, 54*, 1445–1455. https://doi.org/10.1037/dev0000538

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *British Medical Journal, 339*, b2535. https://doi.org/10.1136/bmj.b2535

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86*, 207–236. https://doi.org/10.3102/0034654315582067

Polanin, J. R., & Terzian, M. (2019). A data-sharing agreement helps to increase researchers' willingness to share primary data: results from a randomized controlled trial. *Journal of Clinical Epidemiology, 106*, 60–69. https://doi.org/10.1016/j.jclinepi.2018.10.006

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science, 22*, e12704. https://doi.org/10.1111/desc.12704

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Tricco, A. C., Tetzlaff, J., Sampson, M., Fergusson, D., Cogo, E., Horsley, T., & Moher, D. (2008). Few systematic reviews exist documenting the extent of bias: A systematic review. *Journal of Clinical Epidemiology, 61*, 422–434. https://doi.org/10.1016/j.jclinepi.2007.10.017

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology, 55*, 934–950. https://doi.org/10.1037/dev0000699

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

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

Van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data, 5*, 4. https://doi.org/10.5334/jopd.33

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/

Von Holzen, K., & Bergmann, C. (2018). A meta-analysis of infants' mispronunciation sensitivity development. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1159–1164). Austin, TX: Cognitive Science Society.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 41–56). New York, NY: Russell Sage Foundation.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'* (R package version 1.2.1). Retrieved from https://CRAN.R-project.org/package=tidyverse

History
Received May 20, 2019
Revision received August 26, 2019
Accepted September 19, 2019
Published online March 31, 2020

Open Data
Data and analyses scripts are shared on our Open Science Framework project site (https://osf.io/g6abn/) and on Psych Archives (data: https://www.psycharchives.org/handle/20.500.12034/2185, code: https://www.psycharchives.org/handle/20.500.12034/2186). Analyses have been conducted with the tidyverse (Wickham, 2017) and the metafor (Viechtbauer, 2010) packages in R (R Core Team, 2019).

Funding
This research was supported by grants from 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 Skłodowska-Curie grant No. 659553], the Agence Nationale de la Recherche [ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017].

ORCID
Sho Tsuji
https://orcid.org/0000-0001-9580-4500

Sho Tsuji
International Research Center for Neurointelligence
Institutes for Advanced Studies
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-0033
Japan
tsujish@gmail.com