DOI: 10.1111/infa.12470

RESEARCH ARTICLE

INFANCY THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES WILEY

Ten easy steps to conducting transparent, reproducible meta-analyses for infant researchers

Loretta Gasparini[1,2] | **Sho Tsuji**[3] | **Christina Bergmann**[4]

[1]Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria, Australia

[2]Department of Paediatrics, The University of Melbourne, Royal Children's Hospital, Parkville, Victoria, Australia

[3]International Research Center for Neurointelligence, Institutes for Advanced Studies, The University of Tokyo, Tokyo, Japan

[4]Language Development Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Correspondence
Loretta Gasparini, Royal Children's Hospital, 50 Flemington Rd, Parkville VIC 3052 AUSTRALIA.
Email: loretta.gasparini@mcri.edu.au

Funding information
MEXT WPI research startup fund; Fetzer Franklin Fund; Institute for AI and Beyond, JSPS Grant-in-aid for Specially Promoted Research, Grant/Award Number: 20H05617; JSPS Grant-in-aid for Transformative Research Areas, Grant/Award Number: 20H05919
Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

Abstract

Meta-analyses provide researchers with an overview of the body of evidence in a topic, with quantified estimates of effect sizes and the role of moderators, and weighting studies according to their precision. We provide a guide for conducting a transparent and reproducible meta-analysis in the field of developmental psychology within the framework of the MetaLab platform, in 10 steps: (1) Choose a topic for your meta-analysis, (2) Formulate your research question and specify inclusion criteria, (3) Preregister and document all stages of your meta-analysis, (4) Conduct the literature search, (5) Collect and screen records, (6) Extract data from eligible studies, (7) Read the data into analysis software and compute effect sizes, (8) Visualize your data, (9) Create meta-analytic models to assess the strength of the effect and investigate possible moderators, (10) Write up and promote your meta-analysis. Meta-analyses can inform future studies, through power calculations, by identifying robust methods and exposing research gaps. By adding a new meta-analysis to MetaLab, datasets across multiple topics of developmental psychology can be synthesized, and the dataset can be maintained as a living, community-augmented meta-analysis to which researchers add new data, allowing for a cumulative approach to evidence synthesis.

1 | INTRODUCTION

Infants' cognitive development over the first few years of life shows rapid progress. Developmental researchers are interested in measuring infant development to ultimately develop generalizable theories about the observable phenomena in question. One key tool for this purpose is meta-analysis, a technique to statistically integrate systematically assembled past literature ("The Value of Evidence Synthesis," 2021).

Aggregating over past studies is a promising tool for developmental researchers, who grapple with the specific challenges of conducting infant studies. Since infants are hard to recruit and test compared to adults, infant researchers need to carefully select their experimental or observational groups, often testing a restricted number of conditions or age groups, resulting in small and noisy samples. Low power is an almost inevitable consequence (Bergmann et al., 2018). A robust null effect at earlier ages should become a robust non-null result as infants mature and acquire the skill of interest. Meta-analyses make this possible by increasing sample size and interpolating between studies that test a range of age groups.

With these advantages in mind, we have created MetaLab (metalab.standford.edu), a platform that facilitates conducting and accessing meta-analyses. We provide a step-by-step guide on conducting a meta-analysis in MetaLab, with analyses conducted with the software R and RStudio (R Core Team, 2020; RStudio Team, 2020). This guide can be consulted independently of MetaLab but refers to tools developed within this platform. In this section, we describe the concepts of meta-analysis (i), effect sizes (ii), moderators (iii) and weighting (iv) and explain the value of meta-analyses for informing future research (v). We then provide the step-by-step guide to conducting a meta-analysis.

1.1 | What is a meta-analysis?

To delineate the concepts of a meta-analysis, we define three ways in which a literature can be summarized: narrative (qualitative, literature) review, systematic review, and meta-analysis. A narrative review, involves the authors' selection of literature, oftentimes using unspecified criteria, which they synthesize to support an argument or discuss an issue from their perspective. This is a powerful tool for clarifying concepts and trends in the literature. Systematic reviews are an attempt at an exhaustive overview of the empirical literature pertaining to a specific question. This ideally involves clear, pre-specified, and documented study eligibility criteria. The results are tabulated, and the reviewer discovers what the literature shows, providing a fine-grained and complete overview of the body of empirical evidence. When the focus of a review is a quantifiable and sufficiently homogeneous effect, a meta-analysis can add to a systematic review by using statistical methods to combine the results.

1.2 | Standardized effect sizes

Meta-analyses produce estimates of the effect size of a phenomenon based on all available evidence, which is a more precise measure than can be provided by individual studies. In addition, effect sizes provide a gradual measure, as opposed to the possibly misleading *yes* or *no* dichotomy imposed by *p*-values. To estimate an overall effect size, in meta-analyses, we usually express the outcome of single experiments as standardized effect sizes, how big an effect is, their variance, and how much the results vary. By standardizing effect sizes, we can compare results of different outcome measures on the same

scale, the raw outcomes of which would be difficult to compare, for example, the percentage of trials when a baby looks at the target versus the total looking time.

Effect sizes can be based on differences in group means, binary data (yes/no), or correlations. In this paper, we focus on effect sizes based on (differences in) group means, as this best captures the continuous outcomes of many laboratory-based paradigms. A common such effect size is Cohen's *d*, calculated as the difference in group means divided by the standard deviation (SD) pooled across groups (Borenstein, 2009). The resulting effect size is interpreted such that $d = 0$ indicates no difference between groups. Cohen (1992) proposed that it depends on the construct in question as to what constitutes a theoretically meaningful effect size, but in the case that no such information is available, a rule of thumb is that $d = 0.2$ can be considered a "small" effect size, $d = 0.5$ "medium", and $d = 0.8$ "large". Cohen's *d* has a slight bias to overestimate effect sizes in small samples, so in infant studies it is often adequate to apply a correction resulting in Hedges' *g* (Hedges, 1981; see Section 2.7).

1.3 | Quantifying the effects of moderators

One strength of meta-analyses is the appraisal of moderator variables. Moderators of interest in cognitive development research may relate to the sample (e.g., native language and socio-economic status), or methodology (e.g., testing method and number of trials). While qualitative reviews may suffice for considering a few moderators, they quickly reach their limits as more studies and moderators are included. Meta-analysis also allows for a quantification of moderators' effects, estimating in what direction and with what magnitude moderators change results.

Moreover, meta-analyses can include moderators traversing single studies. One theoretically relevant moderator across developmental studies is age, to assess developmental changes in abilities, behaviors, and responses. A single study might compare babies from a few different age groups with age as a categorical moderator. By virtue of containing more data, meta-analyses have the potential to appraise age as a continuous moderator and shed light on the developmental trajectory of a particular ability or behavior. This can also apply to other moderators of interest.

1.4 | Weighting based on precision

Meta-analysis allows for giving certain studies more weight than others. A commonly used weighting criterion is a measure of precision. Precision can be estimated based on sample size and variability, whereby larger studies and studies with less variability are considered to have higher precision.

These three elements: calculation of effect sizes, quantification of moderators, and weighting, are unique to meta-analyses, and are invaluable for answering a research question considering all available evidence in an objective and reproducible manner.

1.5 | Using meta-analyses to inform later research

Meta-analyses are useful for informing new research. For instance, meta-analysis allows for prospective power calculations (e.g., ManyBabies Consortium, 2020, estimated power based on a MetaLab dataset based on Dunst et al., 2012). Using the effect size and sample size of previous similar studies, one can calculate the likelihood of detecting a true effect in a planned study. Using this information, one can decide how many participants need to be tested to detect an underlying true effect, that is,

power. Meta-analyses can also inform experimental design choices, for instance, the method that has previously led to the largest effect sizes (see Bergmann et al., 2018). Meta-analyses can also be useful for other aspects of experiment planning, such as selecting stimuli (e.g., Rabagliati et al., 2019).

Going beyond "classic" meta-analyses, MetaLab implements the so-called Community-Augmented Meta-Analyses (CAMAs; Cristia et al., 2021; Tsuji et al., 2014), which open meta-analyses to the community. This means a meta-analysis can be updated as new studies emerge or null results are dredged from the file drawer. In this paper, we aim to instruct how to conduct a meta-analysis that is ready to become a CAMA on MetaLab or a similar platform (Burgard et al., 2021). There are various publications that used MetaLab data beyond the initial meta-analysis, demonstrating the added benefit of making such data available (e.g., Bergmann et al., 2017, 2018; Mathur & VanderWeele, 2020, 2021; Tsuji & Cristia, 2017; Tsuji et al., 2020).

2 | TEN STEPS TO CONDUCTING A META-ANALYSIS

Here, we provide a step-by-step guide for conducting a MetaLab-ready transparent, reproducible, and updatable meta-analysis in a topic of developmental psychology. Where relevant we include code for conducting steps in R and RStudio. All code and data files can be found in the Supplementary Materials (osf.io/n9tav/). The code is adapted from MetaLab and previous publications using the MetaLab framework (Black & Bergmann, 2017; Carbajal et al., 2021; Csibra et al., 2016; Gasparini et al., 2021; Rabagliati et al., 2019). We run the data on a simplified dataset on language discrimination from Gasparini et al. (2021). Our code requires the R packages *metagear* (Lajeunesse, 2016), *irr* (Gamer et al., 2019), *metafor* (Viechtbauer, 2010), *tidyverse* (Wickham et al., 2019), *RColorBrewer* (Neuwirth, 2014), *ggplot2* (Wickham, 2016), *gridExtra* (Auguie, 2017), and *MASS* (Venables & Ripley, 2002). You can run the code using the provided data, or on your own data in the MetaLab format (Section 2.6).

2.1 | Choose a topic for your meta-analysis

The topic of your meta-analysis must be clearly defined, and you should be able to justify in your manuscript why it is important that a review is conducted on this topic. The process of deciding on a topic for a meta-analysis is much like deciding on a topic for an experiment. However, because the goal of a meta-analysis is to synthesize the body of literature and come to conclusions based on evidence generalized across many studies, the topics of a meta-analysis will be broader than those of single (series of) experiments. A very high-level topic is "how babies learn language". However, this is too broad for a single meta-analysis. To ensure that a range of studies are appropriate to be synthesized in a meta-analysis, it is fundamental that there is a largely consistent construct of interest being measured across the studies, which would be difficult with a topic at this high level (although Lewis et al., 2016, are conducting a meta-analysis across different language acquisition domains to address questions closer to this level).

A more medium-level topic could be "how babies recognize their native language". Babies have shown evidence for paying more attention to their native language over some foreign languages within days of life. This is thus one piece of the puzzle in answering the high-level question of how babies learn language and the topic that Gasparini et al. (2021) pursued in their meta-analysis. When planning a meta-analysis, it can be useful to identify the seminal paper of a given topic, as it may have presented the issue at this level of scope (e.g., Gasparini et al., 2021, identified Mehler et al., 1988).

Usually, a meta-analysis will address a topic of the medium-level scope, with the goal of estimating the true effect size of a phenomenon across all extant studies.

Later experiments on a given topic may aim to tease out the effects of moderators, and so reduce the scope of their topic to an even lower level, such as "how babies discriminate between languages with different rhythm patterns". If your motivation for conducting a meta-analysis is to assess the effect of moderators, this could be a good reason to choose a more specific topic.

In deciding on a topic for your meta-analysis, you can do a quick scoping review to assess the number and breadth of available studies, and their methodological and sample heterogeneity. Valentine et al. (2010) argue that meta-analysis is the most valid approach to evidence synthesis, so identifying even two studies warrants quantitative synthesis in preference to making inferences based on nonquantitative appraisals, such as counting the number of the significant versus nonsignificant results. So rather than helping you to determine whether to conduct a meta-analysis, considering the number and heterogeneity of available studies will inform how you phrase your research question (see Section 2.2) and assess the generalizability of your findings. You may need to revise your decision along the way, for instance, if you cannot obtain sufficient data about moderators. Make sure you do so transparently and keep track of adjustments.

2.2 | Formulate your research question and specify inclusion criteria

After deciding on your topic, you want to formulate a clear research question, and a list of inclusion criteria. The PRISMA guidelines (Moher et al., 2009) use the mnemonic PICOS for elements that should be included in the research question and inclusion criteria, which we adapt to PECOS for experimental (as opposed to intervention) studies. These elements are the (P) Population being addressed, (E) Experimental condition (the condition of interest), (C) Comparison condition, (O) Outcome (or dependent variable), and (S) Study design. Taking Gasparini et al. (2021) first research question "How do typically-developing infants' abilities to discriminate between languages in the same or different rhythm classes change from birth up to 12 months of age?", see in Table 1 how this question is made up of these PECOS elements (except for Study Design, which was not specified).

Next, specify the inclusion criteria. Consider whether you need to have criteria pertaining to the following details: participants, dependent variable, method, stimuli, research question, and document features (see Table 2 for inclusion and exclusion criteria from Gasparini et al., 2021, fitting into these categories). The goal of these criteria is to arrive at a set of studies that allow you to answer your research question based on all available evidence, but not irrelevant studies.

TABLE 1 PECOS elements of Gasparini et al. (2021) research question

	Gasparini et al. (2021)
Population	Typically developing infants
Experimental condition	Different rhythm classes
Comparison condition	Same rhythm classes
	From birth to 12 months of age
Outcome/dependent variable	Ability to discriminate between languages
Study design	N/A (included any study design)

TABLE 2 Selection criteria from Gasparini et al. (2021)

	Gasparini and colleagues' (2021) selection criteria
Research question	Discrimination or preference between two languages, dialects or accents was the key component of the task
Dependent variable	The dependent variable was a difference in response to stimuli in two different language varieties
Method	Any response measures (e.g., behavioral, neurophysiological) and any test paradigms (e.g., visual fixation, head-turn preference) were considered
Participants	Participants were infants aged from 0 days to 11 months, 31 days Participants were typically developing, born at full-term, with no visual or hearing impairments
Stimuli	Stimuli were derived from continuous, natural speech, presented in the auditory modality. Single sounds, syllables or words, word lists, or backward speech were not eligible. Audio-visual stimuli were allowed if the auditory component fulfilled the above criteria, and the video was consistently included and congruent with the audio. Manipulations of natural speech were allowed (e.g., low-pass filtered or resynthesized speech)
Publication features	The data is not duplicated in the meta-analysis. In the case that the same data is represented in multiple eligible publications, the data from the first peer-reviewed publication were included Any document with unique data was allowed regardless of publication status or type of publication Documents from any years were considered

2.2.1 | Participants

It is important to consider the heterogeneity of the population being tested across studies. Factors like native language(s), bilingualism, neurotypicality, birth complications or sensory impairments may affect results, in which case the factor should either be included as a moderator or specified as an exclusion criterion such that the sample is kept homogeneous along that factor (e.g., excluding preterm babies). If doing so, clarify that results may not generalize to excluded populations.

Consider the eligible ages of participants. Generally, we try to identify a developmental trajectory of the construct of interest, so age will likely be a theoretically important moderator in your meta-analysis. Consider from your content knowledge what stages have been identified as important for the construct under investigation, and what is a suitable age range to synthesize (specify to the level of months in early development, and years for school-aged children and older).

2.2.2 | Dependent variable

Specify the dependent variable, that is, shared between all studies. Is this a within- or between-subjects measure, or are both acceptable?

2.2.3 | Method

Will all methods (e.g., behavioral, gaze, and neurophysiological) and experimental designs (e.g., habituation and familiarization) be considered, or only a subset? Consider whether different methods and designs are expected to yield different results, and whether this is something that can be accounted for in your analyses (see Section Direction of effect sizes), or if it is preferable to have a more homogeneous dataset.

2.2.4 | Stimuli

Specify what stimulus modalities are allowed (e.g., audio, visual, or audiovisual; static or dynamic), what minimal features must be present (e.g., a full word or utterance, a string of three stimuli, a sequence of two actions, a human face), and along what dimensions the stimuli in the experimental and control conditions may vary.

2.2.5 | Research question

To maintain homogeneity in your sample, you might decide that you will only include studies where the aims of the experiment were related to your topic. There may be studies that fulfill all other criteria, but if they were not designed to maximize differences in responses based on the construct you are interested in, inclusion of these studies may add noise to your dataset. If synthesizing across studies that are not asking the same question as you, it might be useful to assess the impact of this decision in a moderator analysis.

2.2.6 | Publication features

Specify whether you will consider documents of any type regardless of publication and peer-review status. Namely, will you also include file-drawer studies, preprints, unpublished theses, or conference presentations? If including unpublished sources, ensure that the same data is not duplicated in your dataset. Consider whether you will include studies conducted or published within a certain period or at any time. And specify whether you will limit the inclusion of data based on language based on practicalities of being able to extract the data. For a meta-analysis to be thorough and overcome publication bias, it should be as inclusive as possible, so exclusions based on the document type should be justified. It can be a viable argument that published studies are likely to contain more high-quality data for reasons including having been written up and undergone thorough analysis. But note that

including unpublished data does not preclude later analyses based on published data only or comparing published and unpublished studies. This set of decisions influences your search strategy, for example, whether you consult scientific search engines only or also launch broad calls for unpublished data (see Section 2.4).

2.3 | Preregister and document all stages of your meta-analysis, to ensure transparency and reproducibility

As it becomes more of a central requirement for single studies (Open Science Collaboration, 2015, 2017), reproducibility is also a topic of increasing importance for meta-analyses. Documentation is key for reproducible meta-analyses where any researcher should be able to come to the same conclusion as the original meta-analyst(s). Throughout this tutorial, we demonstrate how to document processes and decisions to ensure transparency and reproducibility, drawing from our experience and resources that provide detailed guidelines (Laurinavichyute & Vasishth, 2021; Moreau & Gamble, 2020; Page et al., 2021; Polanin et al., 2020). In Table 3, we summarize all details you should include in your protocol, manuscript, and supplementary materials.

It is helpful to preregister your meta-analysis, so that it is transparent which decisions you made before data collection, and which you adjusted based on your findings. You can preregister at multiple stages: before the scoping review, full search, or data analysis (to name a few possibilities). You can preregister either by timestamping your decisions up to this point (topic, research question, criteria, search terms, i.e., ***Sections*** 2.1-2.3) or, more extensively, by writing a protocol, which can be similar in structure to the Introduction and Methods sections of the final report but indicates the *planned* methods before they have been conducted. In this case, you should plan all the processes of *Sections* 2.4-2.10 before starting on them, write and timestamp the protocol, then proceed with the meta-analysis as you described it. PRISMA has published guidelines for what to include in a meta-analysis preregistration (Moher et al., 2015).

Once you have written your preregistration or protocol, you can publish it on an online platform with timestamping and version control. If you want to submit it to a journal as a Stage 1 Registered Report, Center for Open Science (n.d.) provides a list of journals that accept Registered Reports for meta-analyses. You can also publish your protocol immediately, for instance, on Open Science Framework (OSF, Center for Open Science, 2021). Once you can share a protocol and repository of a planned meta-analysis, you are welcome to alert MetaLab (by contacting metalab-project@googlegroups.com) and we will share these on the MetaLab website. This will inform other MetaLab users that a meta-analysis on this topic is ongoing.

Once you start your search and reviewing papers, it is important to document all steps and decisions you make throughout the process. Familiarizing yourself with PRISMA's reporting guidelines (Page et al., 2021) will facilitate writing the final manuscript (***Section*** 2.10). If you need to make any changes to your decisions, document, justify, and timestamp these changes. This can, for instance, be achieved by uploading a new document with tracked changes onto the OSF.

When conducting the statistical analyses, prepare to make your workflows transparent and reproducible, even for users who are less proficient in statistics and the statistical software (see Laurinavichyute & Vasishth, 2021, for guidelines). Include comments in your code describing what you are doing and provide a README text file that guides users in interpreting your file structure

TABLE 3 Summary of what to include in your protocol, manuscript, and supplementary materials

	Protocol and manuscript	Manuscript	Supplementary materials
1. Choose a topic for your meta-analysis	Rationale and objectives		
2. Formulate your research question and specify inclusion criteria	Study eligibility criteria (population, design, dependent variable, publication status, year, language)		
3. Preregister and document all stages of your meta-analysis, to ensure transparency and reproducibility		Funding details, where the review protocol can be accessed, details of any amendments to information provided in the protocol (or where they can be accessed, including justification of amendments and stage of the review process at which the amendment was implemented), where the Screening Decision Spreadsheet, meta-dataset, analytic code, any other materials can be accessed (e.g., link to public repository, or contact details of author responsible for sharing materials upon request)	
4. Conduct the literature search	All information sources (databases, registers, websites, organizations, mailing lists, reference lists, experts, and other sources), exact search terms/syntax, number of results screened (e.g., if not all from a database search)	The date when each source was searched or consulted, the number of studies yielded by each search source (e.g., hits of a database search)	
5. Collect and screen records	Title, abstract, and full-text screening, and selection processes (including double screening)	PRISMA flow-chart, 'near-misses' with justifications of why they were excluded. Inter-rater reliability, where and why disagreements arose, and how they were resolved.	Document all decisions in the Screening Decision Spreadsheet, and make it publicly available and findable

TABLE 3 (Continued)

	Protocol and manuscript	Manuscript	Supplementary materials
6. Extract data from eligible studies	Data extraction and coding processes, all variables for which data were sought	Characteristics of included studies	Keep all up-to-date data in the Data Spreadsheet, clearly describe all columns and their specifications in the Codebook and make both publicly available and findable. When you publish your meta-analysis, version-control the data you used in the published version of the analyses, so that any changes to the dataset (correction of errors or new data) are identifiable
7. Read the data into analysis software and compute effect sizes	Software, packages and version numbers, the type of effect size used for synthesis of results and how it is calculated or estimated, processes for deciding eligibility for synthesis, methods for data cleaning and preparing for synthesis (dealing with missing data or conversions)	Characteristics of all included studies, and studies included in all subgroup analyses	Keep a reproducible pipeline of all code (so that someone can download the data and code, and run all the code and obtain the same results), write a user-friendly code (annotations describing each process, give meaningful names to objects), and make all data and code publicly available and findable
8. Visualize your data	Software, packages and version numbers, methods for assessing risk of bias	Results of individual studies (forest plot), risk of (publication) bias assessment (funnel plot)	Keep a reproducible pipeline of all code, write user-friendly code, and make all data and code publicly available and findable
9. Create meta-analytic models	Software, packages and version numbers, type of meta-analytic model (e.g., random effects multivariate), weighting (e.g., inverse-variance), methods for quantifying statistical heterogeneity, the between-study variance estimator used (e.g., REML), methods for exploring heterogeneity (e.g., subgroup analysis, meta-regression, and if so, which factors are explored), sensitivity analyses	For all analyses (including sensitivity, subgroup, and moderator): Summary estimates, direction of effects, precision (SE or 95% CI), *p*-values, statistical heterogeneity; the robustness of main analyses given sensitivity analyses	Keep a reproducible pipeline of all code, write a user-friendly code, and make all data and code publicly available and findable

(Continues)

TABLE 3 (Continued)

	Protocol and manuscript	Manuscript	Supplementary materials
10. Write up and promote your meta-analysis	How you will assess certainty in the body of evidence	Overall level of certainty in the body of evidence, a general interpretation of the results, limitations of the evidence, limitations of the review processes used, implications of the results (e.g., for practice and policy), recommendations for future research	

and reproducing the analyses. In RStudio, you can create RMarkdown files that produce the code output with your comments so the reader can view the results without rerunning the code, or they can open the file in RStudio and run the file to achieve the same results you reported. Cite the packages that you used and provide session information (e.g., using the *citation() and SessionInfo()* functions in R), since factors like software and package version can alter results even when run with the same code.

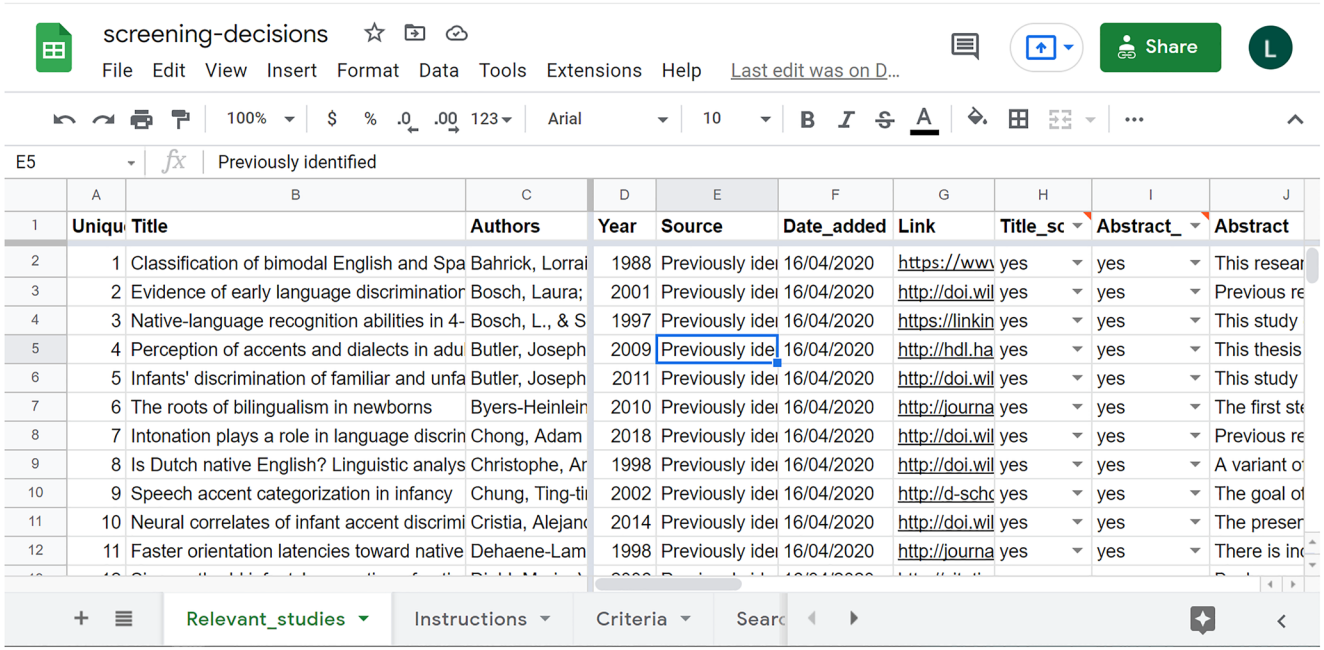
2.4 | Conduct the literature search

Studies can be aggregated in different ways: (i) studies you and your co-authors previously identified, (ii) through database search(es), (iii) by searching for studies that cite a seminal paper (forward citation search), (iv) by scanning the references of eligible studies (backward citation search), and (v) by consulting experts. Ideally, you will combine all these methods to maximize the likelihood of capturing all available literature.

MetaLab provides a Screening Decision Spreadsheet template (see Figure 1 and Supplementary Materials) for entering all studies that arise from the search, to document your screening for eligibility (see Section 2.5). It is useful to also document your literature search methods under the tabs "Instructions", "Criteria", "Search protocols", and "Notes". Below we refer to tab and column names from the MetaLab Screening Decision Spreadsheet, but the same principles apply if using a different template.

2.4.1 | Previously identified studies

Add all the studies you and your co-authors are already aware of that fit the inclusion criteria, through your knowledge of the topic and that you came across in your scoping phase.



	A	B	C	D	E	F	G	H	I	J
1	Uniqu	Title	Authors	Year	Source	Date_added	Link	Title_sc	Abstract_	Abstract
2	1	Classification of bimodal English and Spa	Bahrick, Lorrai	1988	Previously ide	16/04/2020	https://ww	yes	yes	This resear
3	2	Evidence of early language discriminatior	Bosch, Laura;	2001	Previously ide	16/04/2020	http://doi.wil	yes	yes	Previous re
4	3	Native-language recognition abilities in 4-	Bosch, L., & S	1997	Previously ide	16/04/2020	https://linkin	yes	yes	This study
5	4	Perception of accents and dialects in adu	Butler, Joseph	2009	Previously ide	16/04/2020	http://hdl.ha	yes	yes	This thesis
6	5	Infants' discrimination of familiar and unfa	Butler, Joseph	2011	Previously ide	16/04/2020	http://doi.wil	yes	yes	This study
7	6	The roots of bilingualism in newborns	Byers-Heinlein	2010	Previously ide	16/04/2020	http://journa	yes	yes	The first ste
8	7	Intonation plays a role in language discrin	Chong, Adam	2018	Previously ide	16/04/2020	http://doi.wil	yes	yes	Previous re
9	8	Is Dutch native English? Linguistic analys	Christophe, Ar	1998	Previously ide	16/04/2020	http://doi.wil	yes	yes	A variant of
10	9	Speech accent categorization in infancy	Chung, Ting-ti	2002	Previously ide	16/04/2020	http://d-scho	yes	yes	The goal of
11	10	Neural correlates of infant accent discrimi	Cristia, Alejanc	2014	Previously ide	16/04/2020	http://doi.wil	yes	yes	The preser
12	11	Faster orientation latencies toward native	Dehaene-Lam	1998	Previously ide	16/04/2020	http://journa	yes	yes	There is inc

FIGURE 1 MetaLab screening decision spreadsheet

2.4.2 | Database search

Decide in which database(s) you will search. Many MetaLab meta-analyses have used Google Scholar as it is not behind a paywall and has wide coverage, but it does have downsides including not being reproducible and being biased by previous search terms. A Google Scholar search is better suited as a supplementary resource, combined with at least one reproducible database (Gusenbauer & Haddaway, 2020) such as PubMed, Scopus, PsycINFO, or Web of Science. You can also search preprint servers like PsyArXiv.

Google Scholar does not include functions for documenting and exporting literature searches that other databases possess. A useful program for conducting a Google Scholar search is Harzing's Publish or Perish (Harzing, 2007), from which you can export searches to a .csv file.

To start defining possible database search terms during the scoping phase of your project, you can draw from your research question, including the phenomenon of interest (e.g., "language discrimination"), the participant group (e.g., "infant", "child"), and the method (e.g., "looking"). By entering these terms into the database, you may find that the search yields thousands of results, so you could limit your search to just the first 500 or 1000 results. You could then conduct more narrow searches by trying additional search terms. Keywords of relevant articles can be a good source for this. Decide upon search terms and limits so that it will be feasible to screen all records, but so that it can be considered a thorough and exhaustive search of the literature. Consulting with a librarian can help ensure your search terms and database choices are appropriate.

To finalize your search terms, you can first use the above strategies to narrow down a realistic search strategy, preregister this protocol, and then conduct the actual search. In your report, document the search terms and dates of the searches, number of results the search yielded, and how many were scanned.

2.4.3 | Forward citation search

A forward citation search takes a seminal study and screens relevant papers citing this study. You can do this, for instance, by using the same search terms as your database search but only including studies citing the seminal study. Make sure you document the same details as any other database search.

2.4.4 | Backward citation search

You could also screen the citations of eligible studies or any review papers on the topic to capture any studies missed by the previous methods. The processes of screening reference lists and consulting experts (see Section 2.4.5) starts after you have started the record collection and screening process (***Section*** 2.5), so these steps will overlap and be iterative as new eligible studies arise. As you do not have control over when new articles are published or authors will respond to your emails or call for studies, you could indicate in your preregistration that for your initial analyses, you will include data from any studies you become aware of and can access up until a certain date. If you become aware of additional studies beyond this date, you should keep a record of them for later addition, especially if you aim for your meta-analysis to be a living database following the CAMA idea (Tsuji et al., 2014).

2.4.5 | Expert consultation

Contacting experts in the field will bring your attention to studies missed by your literature search or screening, as well as unpublished (in progress or file-drawer) data. This helps to overcome publication bias. One way to contact authors is by posting calls for studies in mailing lists that are topically relevant for your meta-analysis. Define in your preregistration which mailing lists you will post to. We provide in the Supplementary Materials a template call for studies.

You can also ask authors of studies that are included in your meta-analysis if they know of studies you have missed or if they have any unpublished data. You will likely be contacting authors for additional details of their study (see Section 2.6) and can ask this at the same time (see Supplementary Materials for a template for this email). If you can, define in your preregistration the criteria for author contact. Keep track of the authors you contact under the "Authors_contacted" tab of the MetaLab Screening Decision Spreadsheet, as well as additional experts you are yet to contact (or decide not to contact, indicating reasons). This record will be important once you report where you retrieved studies from as required in the PRISMA protocol. To increase transparency, try to be as detailed as possible, noting when an author replied and what information they provided.

2.5 | Collect and screen records

The records elicited by the previous step's literature search methods should be added to your Screening Decision Spreadsheet (see Figure 1) as you identify them. Specify in the column "Source" where you first came across the record (e.g., "Previously identified, "Google Scholar search 1"), and in "Date_added", the date the record was identified.

Once you have conducted your (first) search, you can start screening titles and entering your eligibility decision in "Title_screening_decision" and identifying yourself as "Title_screener". To remove duplicates, order the rows alphabetically by title. From the title you can often tell if a study is investigating a different topic than yours or if the participant group is not eligible, but when in doubt, retain studies for further screening. For studies that passed the title screening, retrieve the abstract (or indicate if it cannot be retrieved) and paste it into the column "Abstract". Identify yourself as "Abstract_screener" and after reading the abstract indicate "Abstract_screening_decision1". The abstract should clarify the study topic and participant group and may mention the stimuli. Add a column for each inclusion criterion and fill them with *yes* or *no* to track which records fulfilled which criteria. Tracking the reason(s) for exclusion can be useful if you wish to conduct a second search on a related topic.

For all records that passed abstract screening, retrieve the full text (or indicate if it cannot be retrieved). Identify yourself as the "Fulltext_reviewer". Scan the full text; the end of the Background (research questions) and the Methods (participants, stimuli, measures, and design) will indicate whether the study fulfills your inclusion criteria.

Following PRISMA reporting standards, you should use the PRISMA flow diagram to describe the results of your search and selection processes. You can fill in the template provided on the PRISMA website (PRISMA, 2021). To obtain the number of records for each stage, see the "PRISMA_flow_diagram" tab of the MetaLab Screening Decision Spreadsheet. This tab tells you how to set filter settings in the "Relevant_studies" tab, and the number of rows in each filter setting is the number to enter in each stage of the PRISMA flow diagram.

2.5.1 | Double screening

Double screening is useful to ensure multiple screeners interpret the inclusion criteria the same way and identify any possible sources of disagreement. It also prevents human error. Decide in advance and indicate in your preregistration how many studies will be double screened, for instance 10% of all abstracts deemed relevant from the titles. The R package *metagear* (Lajeunesse, 2016) can be used to randomly assign a certain percentage of abstracts to different co-authors, and provides an interface for them to screen abstracts. In the Supplementary Materials, we provide a file called *screening_assignment.Rmd* demonstrating this process.

The R package *irr* (Gamer et al., 2019) can be used to calculate Cohen's *kappa* for inter-rater reliability analysis (McHugh, 2012). We provide the code in the Supplementary Materials, in the file called *screening_irr.Rmd*. Indicate where and why disagreements arose between the screeners, and how they were resolved. If agreement is low, consider (and document) whether this indicates that the eligibility criteria are unclear and need to be refined, and how to proceed.

2.6 | Extract data from eligible studies

After screening, you are ready to start appraising the content for quantitative synthesis. There are five main groups of variables you should code: (i) publication descriptors, (ii) domain-specific methodological variables, (iii) domain-specific participant variables, (iv) topic-specific variables, and (v) data for calculating effect sizes. We describe their conceptual relevance, with more detailed descriptions about how to fill in each column on the MetaLab website.

MetaLab provides a Data Spreadsheet template (see Supplementary Materials) to help you extract the data from studies that are needed to make your dataset compatible with MetaLab. In addition to these mandatory fields, you can add as many columns as you need for further variables. Make a copy of this Google sheet, rename it, and add it to your repository so you can start adding the data.

Some of the columns in the template can only be filled with a specific data type, for example, numbers. This is so the information can be automatically added to MetaLab. The Codebook in the template assists you in keeping track of the mandatory fields and you should explain any optional fields you add so that anyone else can make sense of the dataset. The Data validator application on MetaLab allows you check whether your data is compatible with the MetaLab structure.

MetaLab currently contains mainly behavioral experimental data. You should decide in ***Section*** 2.2 whether it is worthwhile also including neurophysiological data in your meta-analysis. If you decide to include neurophysiological data, consider how you will code these data, which are often multidimensional and therefore do not readily map onto MetaLab's response categories. Will you add a new row for each channel and location, or only one global average for each group, and how will you calculate this? Note that this is a complex issue, and MetaLab does not currently have a "best practice" solution, however, you can look at previous meta-analyses including neurophysiological data to guide you (e.g., Tsuji & Cristia, 2014).

2.6.1 | Publication descriptors

Several columns describe a record: study identifiers, peer-review status, and who added the data. There will likely be studies that contribute many rows of data, through multiple experiments and different conditions, as well as cases where the same participants will have contributed data to more

than one experiment within a study. Thus, we provide columns for indicating these cases so that covariance arising from repeated measures can be considered during the analysis (see Section 2.9).

2.6.2 | Domain-specific methodological variables

The next set of columns concern the methods. These are standardized in MetaLab, to ensure compatibility with the analysis tools and facilitate the synthesis of many datasets. The column *method* refers to the paradigm that was used, such as head-turn preference procedure (HPP) or central fixation (CF). The response mode is (partially) determined by the methods, such that the method HPP involves a behavioral response, while CF is oculomotor. But for both, the dependent measure is looking times, while a study using electroencephalogram (EEG) measures amplitude or latency of brain responses. The full list of possibilities can be found in the Codebook. If you need to add your own categories, first ensure they are not already captured, and if not, enter the details of the new level so that future users will understand your data.

2.6.3 | Domain-specific participant variables

To facilitate moderator analysis, certain common variables are included as compulsory columns. These include the group of participant's mean age (in days), sex distribution (expressed as proportion of girls), native language, and whether they come from a specific population like bilinguals or babies born preterm.

2.6.4 | Topic-specific variables

This group of variables depends on what other factors may moderate the effect. Some would be obvious from factors you mention in your research question and hypotheses. Others could be control variables relating to the stimuli or design that many studies in your review mention.

2.6.5 | Data for calculating effect sizes

Here, we focus on one type of standardized mean difference effect sizes, namely Hedges' g (see Section 1.2). To calculate effect sizes, you will need to specify the participant design (see Section Participant design) and add the number of participants. Then, you will need one of either (i) the means and SD, (ii) the t-value and correlation between measures in within-participant designs (see Section Correlations between measures in within-participant designs), or (iii) the F-value and correlation. If you have access to more than just the minimum details (e.g., the means and SDs as well as a t-value), it is useful to report all the information you have in the spreadsheet; it can help you check for your own or the original authors' reporting errors (e.g., a small difference in means should not yield a large t-value). Note that if reporting an F-value, it must be the value of the main effect of interest, not the F-value of an interaction. See Lipsey and Wilson (2001) for more on how to calculate effect sizes from F-values.

Participant design

To calculate effect sizes, you will need to specify the participant design, that is, whether two groups of participants were compared to each other on the outcome measure (between design), one group's performance was compared in two different conditions (denoted a "within-two" design in the MetaLab Data Spreadsheet), or one group's performance was coded as one measure such as percent correct, a difference score, or a score compared to chance level ("within-one").

Direction of effect sizes

As a reminder, Cohen's *d* or Hedges' *g* effect sizes are calculated as the difference in group or condition means divided by the pooled standard deviation. MetaLab calculates effect sizes as the mean of condition 1 (x_1) minus mean of condition 2 (x_2) and it is important to be consistent in what x_1 and x_2 refer to so that the direction of effect sizes is interpretable. Choose a coding direction that makes sense in terms of your predictions and general trends in the data. In MetaLab, we tend to code data such that a familiarity effect (greater response to more familiar or naturalistic stimuli, or the habituated or familiarized condition) is coded as a positive effect size (where $x_1 > x_2$, and $t > 0$), and a novelty effect is coded as a negative effect size ($x_2 > x_1$, and $t < 0$), so you are welcome to follow this guideline.

Note that when it comes to measures of latency, a shorter response (e.g., a faster head turn or neurophysiological response) indicates a stronger effect. Make sure you account for this, either by switching which condition is entered into the columns for x_1 and x_2, or by writing code that flips the direction of latency effect sizes, so that the direction of preference is consistent throughout your whole dataset. In the Supplementary Materials, we flip the direction of effect sizes in the code, so if you use this code, you should enter the conditions into the x_1 and x_2 columns consistently with all other response modes.

Correlations between measures in within-participant designs

A *t*- or *F*-value calculated in a "within-two" design takes into account the within-subjects correlation of raw results, therefore in order to calculate effect sizes from a *t*- or *F*-value you need to also include a value specifying the correlation in participants' responses to each condition (Morris & DeShon, 2002). This correlation value is rarely reported, so the code that MetaLab uses estimates or imputes missing correlations in other ways. The more correlation values that are included in the dataset, the more accurate the imputation process will be. For this reason, it is helpful to report *t*- and *F*-values whenever you can obtain them even if you already have the means and SDs. Likewise, if you can obtain raw data from an author, you will be able to calculate the means and SDs, but it is also helpful to calculate the correlation between each participant's response to both conditions. You should report in your manuscript how many exact correlation coefficients you could calculate, how many you estimated or imputed, and how you did so.

2.6.6 | Data extraction tools and guidelines

Here, we provide tips on extracting and obtaining data from studies. Many relevant details are included in the original publication, so it is useful to have a clean copy of the paper where you visually highlight each piece of information you obtain from the paper to add to the Data Spreadsheet. This makes it quicker to go back and check where you found that information.

In some cases, the exact means and SDs will not be reported but they will be shown in a figure. You can use WebPlotDigitizer software (Rohatgi, 2021; in-development R package *digitizeR*; Rohatgi, 2020) to estimate values from a figure. Note that this is not entirely accurate, especially in

lower quality images, so it is always preferable to try to obtain the exact values from authors. Specify if you estimate values from figures (in the column *source_of_data*). Note that many figures show standard error (SE) bars *not* SD.

After extracting all data from the publication, some data may still be missing. Using the template in Supplementary Materials, try to contact the author to ask for missing details. Ask your collaborators if they know any of the authors you need to contact, as an email from an acquaintance may be more effective. Generally, you can try to contact the corresponding author, and then the first or last author if you do not get a response and email one follow-up email about 2 weeks after getting no response. If you cannot obtain the necessary information, note this in your PRISMA flowchart and keep track of your (attempted) correspondence with co-authors in the Screening Decision Spreadsheet.

Now you are ready to send your dataset to MetaLab, pending your co-authors' permission and any embargo on your data depending on where you want to publish them. Even if you have not started statistical analyses, you can contact us at metalab-project@googlegroups.com and we can add your dataset so that viewers of the website can view the data, see you listed as a curator, and use the visualization and power calculation tools on your dataset. Note that publishing your meta-analysis on MetaLab does not preclude later publication in a journal paper (e.g., Von Holzen & Bergmann's, 2021, preliminary dataset was added to MetaLab before publication).

2.7 | Read the data into analysis software and compute effect sizes

In the Supplementary Materials, we provide the file *data_analysis.Rmd* to read the file *dat.csv* (a subset of data from Gasparini et al., 2021) into R and calculate effect sizes. Note that datasets uploaded onto MetaLab undergo the same steps automatically before displaying on the website. In this code for calculating effect sizes, Cohen's *d* is calculated from means and SDs (Lipsey & Wilson, 2001), otherwise from *t*- or *F*-values and correlation coefficients (Dunlap et al., 1996), then Cohen's *d* is corrected to Hedges' *g* (Morris, 2000).

2.8 | Visualize your data

The file *data_analysis.Rmd* contains code for creating visualizations using *ggplot2* (Wickham, 2016). In practice, you may visualize your data alongside running models, although we report them as separate steps for simplicity. Visualizing the data before running models can familiarize you with the data and alert you to outliers, prompting you to check for errors. We review here the most common visualizations: forest, bubble, box, and funnel plots. Forest plots give an overview of all the calculated effect sizes in your dataset and their 95% CIs (see Figure 2), with the option to group by variables of interest (e.g., by method in ManyBabies Consortium, 2020).

A bubble plot lets you visualize effect sizes by a continuous variable, such as age, where the size of individual datapoints represents another factor. Figure 3 shows all effect sizes from *dat.csv* by age in days, with the size of the circles representing inverse-variance (so that more precise effect sizes are larger). We can also color- and shape-code the effect sizes by a categorical variable, such as in Figure 4, which visualizes peer review status.

You can create a box-plot of effect sizes by method to visualize the impact of categorical moderators (Figure 5, code adapted from STHDA, n.d.). This plot illustrates that CF yields more negative effect sizes than the baseline HPP, which turns out to be a significant difference in the model output (see Section 2.9).

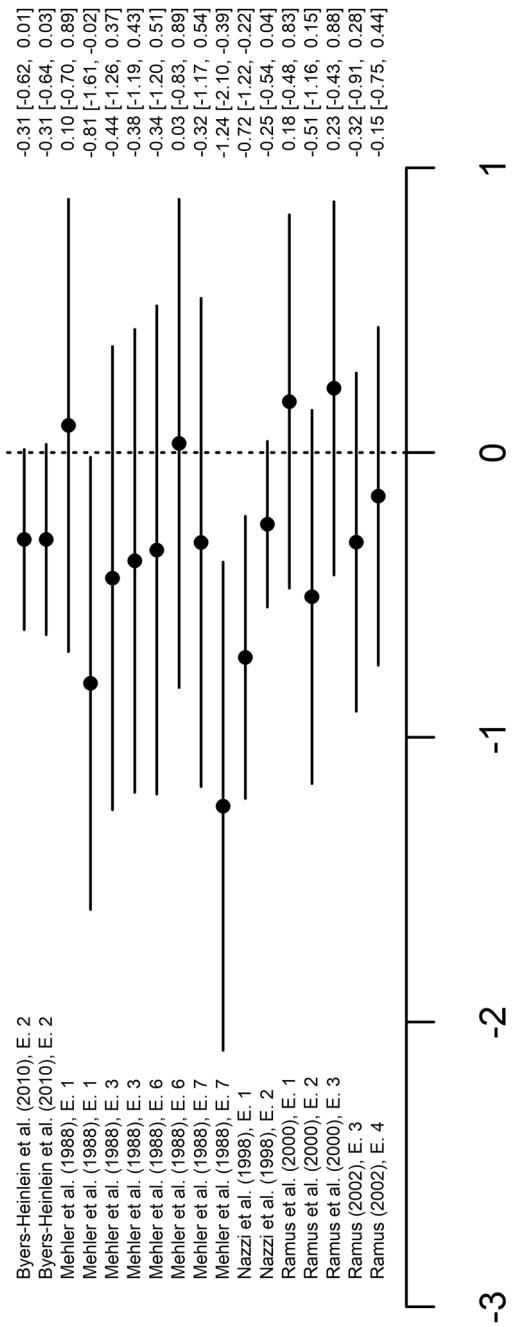


FIGURE 2 Forest plot of effect sizes and 95% CIs (*x*-axis, and *y*-axis, right) by study experiment (*y*-axis, left). Data subset from Gasparini et al. (2021), including only participants under 2 months of age

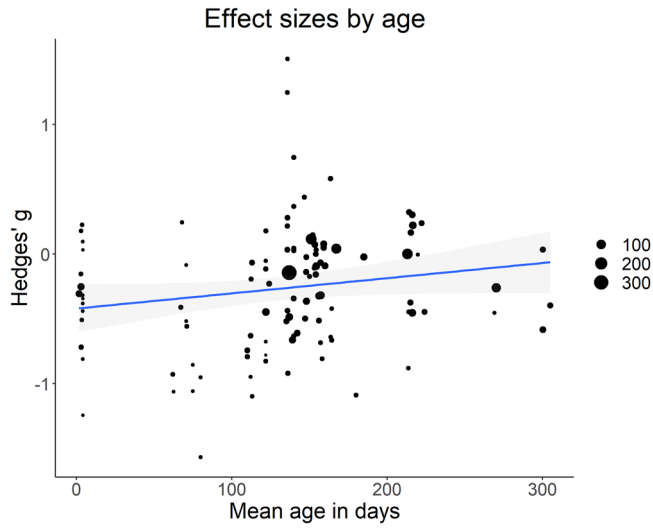


FIGURE 3 Effect sizes (*y*-axis; negative values indicate novelty effect) by age in days (*x*-axis). Bubble sizes represent weighing by inverse of standard error whereby larger bubbles indicate greater precision. Linear fit suggests that effect sizes move toward the null with age, but 95% CI shows substantial overlap. Data subset from Gasparini et al. (2021)

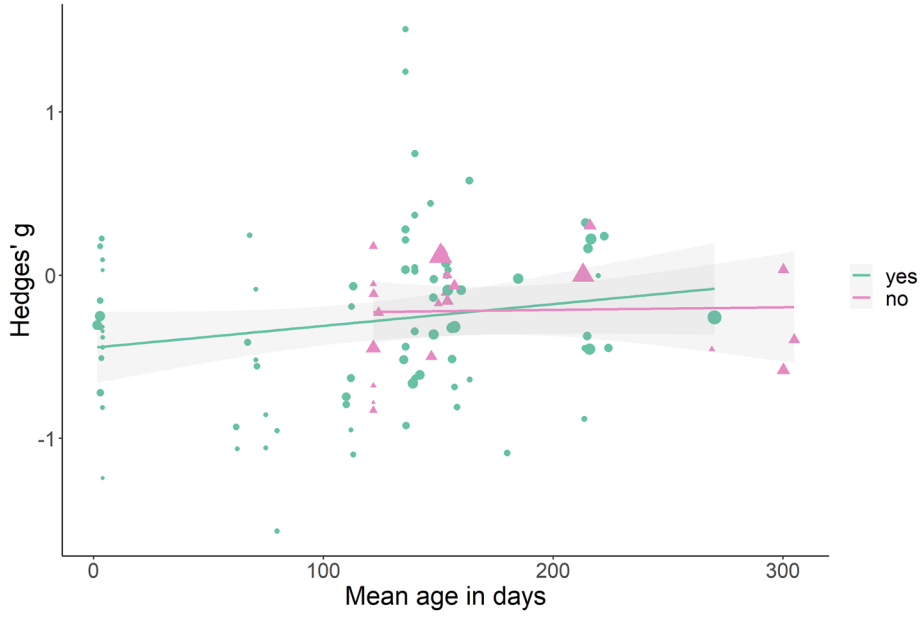


FIGURE 4 Effect sizes (*y*-axis; negative values indicate novelty effect) by age in days (*x*-axis). Color- and shape-coded by peer-review status (yes: green circles, no: pink triangles) and weighted by inverse of standard error whereby larger bubbles indicate greater precision. Linear fit suggest effect of age may be attenuated in the non-peer-reviewed subset. Data subset from Gasparini et al. (2021)

Funnel plots are used to assess the risk of bias in your meta-data. Funnel plots show each effect size along the *x*-axis by SE (an approximation of precision) on the *y*-axis, with a funnel in white spreading ±1.96 SE centering around the calculated overall effect size to illustrate a 95% CI (Figure 6).

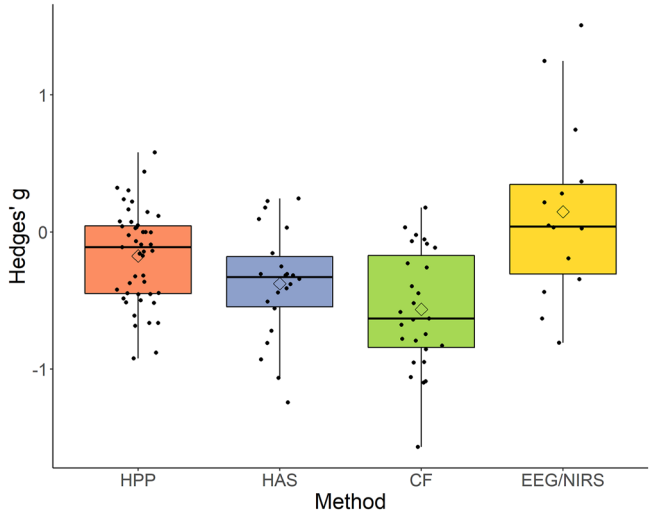


FIGURE 5 Boxplot of effect sizes (*y*-axis; negative values indicate novelty effect) by method (*x*-axis). Diamonds indicate mean values. It suggests that CF method yields greatest novelty effects. Data subset from Gasparini et al. (2021). CF, central fixation; EEG/NIRS, neurophysiological; HAS, high-amplitude sucking; HPP, head-turn preference procedure

In Figure 6, datapoints are color-coded by method and peer-review status. In the case of no sample heterogeneity or publication bias, datapoints would evenly cluster around the overall estimated mean within the funnel, converging closer to this mean the higher the precision. Asymmetry around the mean suggests bias in reporting results in a certain direction (Sterne et al., 2011). You can also create a contour-enhanced funnel plot that centers around zero, the value under the null hypothesis of no effect, with contours of different shades corresponding to different levels of significance (Figure 7). Run Egger's test for funnel plot asymmetry, to see if there is significant asymmetry (if $p < 0.05$). Figures 6 and 7 show a substantial amount of heterogeneity and asymmetry. See Rabagliati et al. (2019) for an example funnel plot that more clearly centers around the mean effect size. Appraising the likelihood of publication bias may inform how you interpret the strength of the evidence yielded from your results.

2.9 | Create meta-analytic models

The file *data_analysis.Rmd* continues with code for running meta-analytic models. Using the *metafor* package (Viechtbauer, 2010), you will run inverse-variance-weighted random effects multivariate meta-analytic regression models. *Multivariate meta-analytic regression* means that the effect sizes calculated from individual studies are regressed against multiple variables of interest. A *random effects* model assumes that the true underlying effect being investigated can vary between studies. By adding nested random effects, we additionally allow the model to account for shared covariance between effect sizes coming from the same experiment or study or involving the same participants. MetaLab's approach for accounting for shared variance is to include in all models the random effects of experiment nested in participant, and participant nested in study. *Inverse-variance-weighted* describes our method of giving more weight to more precise effect sizes according to their smaller variance, which comes about from more precise measurement tools and larger sample sizes (see Section 1.4).

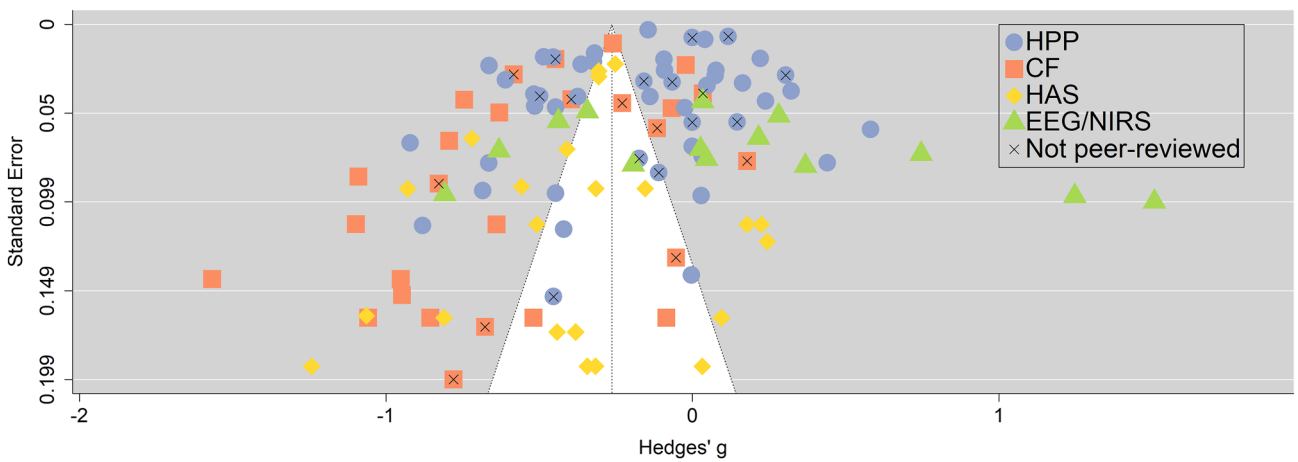


FIGURE 6 Funnel plot of effect sizes, centered around the estimated overall effect size. Color- and shape-coded by method, and cross indicates a record was not peer-reviewed. Datapoints are asymmetrical and heterogeneous which warrants moderator analysis and may attenuate confidence in the overall effect. Data subset from Gasparini et al. (2021)

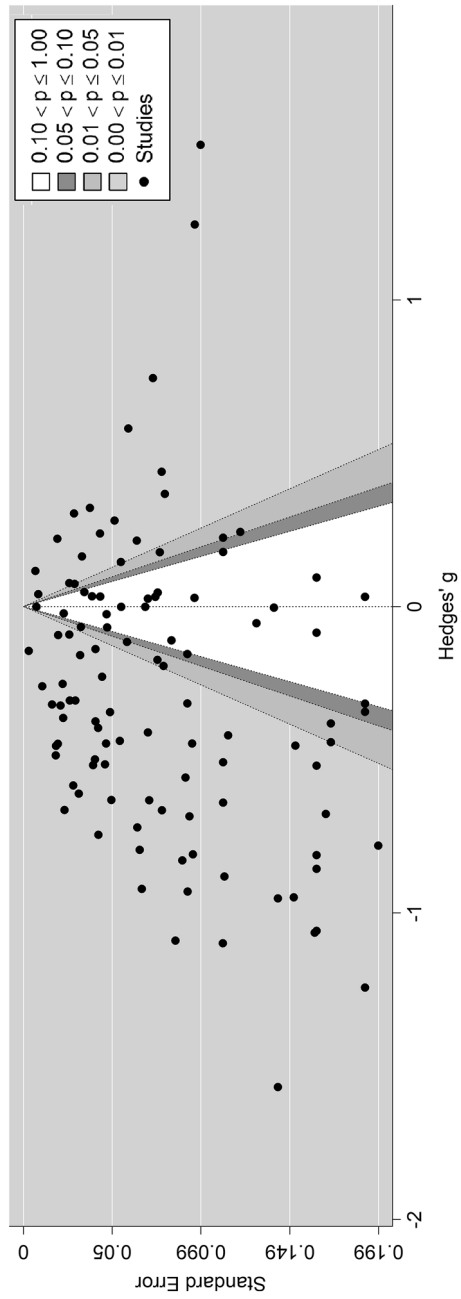


FIGURE 7 Funnel plot of effect sizes centered around zero, the value under the null hypothesis of no effect. Background shade represents level of significance that an effect size is significantly different from zero. It shows many significant novelty effects. Data subset from Gasparini et al. (2021)

Start with a model without any moderators to calculate an overall estimated effect size for the phenomenon you are studying. Table 4 shows model output from the Supplementary Materials, where we see the estimated overall effect size is $g = -0.27$ (95% CI [−0.38, −0.17], $p < 0.0001$). Next, depending on your research questions, you will run models with moderators. In *data_analysis.Rmd*, we use the MetaLab variables 'method' and 'peer_reviewed' as examples.

Consider how you want to contrast code your moderator variables, as this will determine what the intercept and slopes of your model refer to (see Schad et al., 2020). For example, for any categorical factor with two levels (e.g., 'peer_reviewed' yes/no) you could use successive difference coding (using the *MASS* package, Venables & Ripley, 2002) which means that the intercept will indicate the grand mean (see Table 5 for an example of how to report this contrast coding and Supplementary Materials for the R code to set this). For categorical factors with more than two levels (e.g., 'method', HPP/HAS/CF, etc.), you could use simple contrast coding where you set the baseline to be one of the levels (e.g., the most common level, or the one closest to zero, see Table 5).

Continuous factors (e.g., 'mean_age') are often best z-scaled and centered, so that the grand mean shows the estimated effect size for when the continuous factors are at their mean, and the slopes indicate changes in effect sizes per 1 SD increase of the continuous factor. This is particularly the case if you have multiple continuous factors in the same model as it can introduce collinearity (Afshartous & Preston, 2011).

In the Supplementary Materials, we run through an example where we start with a maximal model with all the moderators we are interested in, then run a model comparison of the maximal model with one factor, method, excluded. This full model approach means that the moderator estimate in the model output represents the estimated effect of the moderator when all other included variables are of the same value or level. Depending on your research question, you might have good reasons to report minimal models that do not adjust for other variables (see VanderWeele, 2019; Wysocki et al., 2020, on confounder selection).

Specify in your protocol what criterion you will use for identifying the "best-fitting" model, such as the model with the lowest Akaike Information Criterion (AIC; Akaike, 1974), or if the results of a Likelihood Ratio Tests (LRT) comparing the reduced and full model is significant at the level of $p < 0.05$. Then, look at the model output of the best-fitting model to see the direction and magnitude of effects. In our example, we find that the effect of method is significant (LRT = 11.46, $p = 0.010$) whereby the estimated difference between the baseline HPP and CF in infants of the same age and publications of the same peer-review status is −0.35 (95% CI [−0.57, −0.13]; see Table 6).

2.10 | Write up and promote your meta-analysis

Finally, you are ready to write up your Results and Discussion. Follow PRISMA guidelines (Page et al., 2021) for what to include in your manuscript. Also helpful is the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach of assessing confidence in cumulative evidence (Guyatt et al., 2011).

TABLE 4 Model output of meta-analytic model showing estimated overall effect size (data subset from Gasparini et al., 2021)

Estimate	SE	95% CI	z	p
−0.2729	0.0523	[−0.3753, −0.1704]	−5.2186	<0.0001

TABLE 5 Moderating factors, levels and contrast coding from Gasparini et al. (2021)

Factor	No. levels (*k*)	Levels	Contrast coding
Peer-reviewed	2		Successive difference (baseline = grand mean)
		Yes	−0.5
		No	0.5
Method	4	HPP	Simple (baseline = HPP)
		HAS	−0.25
		CF	−0.25
		EEG/NIRS	−0.25

TABLE 6 Model output of meta-analytic model showing effects of moderators (data subset from Gasparini et al., 2021)

	Estimate	SE	95% CI	*z*	*p*
Intercept	−0.2010	0.0685	[−0.3352, −0.0668]	−2.9354	0.0033
Method (HAS)	−0.0373	0.1717	[−0.3738, 0.2992]	−0.2172	0.8280
Method (CF)	−0.3482	0.1127	[−0.5691, −0.1272]	−3.0884	0.0020
Method (EEG/NIRS)	0.1660	0.1912	[−0.2088, 0.5408]	0.8679	0.3855
Peer-reviewed (no-yes)	0.1951	0.1171	[−0.0345, 0.4246]	1.6657	0.0958
Mean age (scaled)	0.0420	0.0609	[−0.0774, 0.1613]	0.6888	0.4910

Note: Intercept reflects estimated grand mean for babies aged 4.16 months using HPP method. Slopes for "Method" indicate the estimated average effect size difference between the given method and HPP for participants of the same age and publications of the same peer-review status. Slope for "Peer-reviewed" indicates the estimated average effect size difference between non- and peer-reviewed publications when participants are the same age and the same method is used. Slope for "Mean age (scaled)" indicates the estimated average effect size difference for every additional SD (2.4 months) in age when the same method is used and peer-review status is the same.

Add your data, analysis code, and the preprint of your manuscript to a repository (pending the requirements of the journal you wish to submit to) and promote your meta-analysis! The authors of original studies will likely be interested, especially those who replied to your requests for data and study details, and who you acknowledge in your manuscript. Remember to share your data with MetaLab (contact metalab-project@googlegroups.com) so users can view your dataset, find your manuscript when it is ready to share, and see you listed as a curator so that you can be informed of new data to be added to your dataset.

3 | LIMITATIONS

MetaLab is a work in progress, and we welcome suggestions and collaborations. We are working on an R package, *metalabr* (Iverson et al., 2021), that will streamline ***Section*** 2.7-2.9 in this tutorial. Some other features we have not covered here are risk of bias in individual studies (Sterne et al., 2019), power analysis of meta-analysis (Griffin, 2020), systematic coding of neurophysiological data following the Brain Imaging Data Structure (BIDS; Gorgolewski et al., 2016), and other types of effect sizes, namely correlation and odds ratio (Borenstein, 2009). MetaLab is open to proposals for better accommodating different types of data. We note also that meta-analyses are not without their own limitations and their results should be presented accordingly (Lewis et al., 2020).

4 | CONCLUSION

In this paper, we have provided a step-by-step tutorial on how to conduct a meta-analysis, with the objective of lowering the hurdles to use this powerful tool of cumulative science. With your meta-analytic dataset, you will be able to provide researchers with an overview of the body of evidence in this topic in its current form. This can inform future researchers by allowing for power calculations to decide on sample size, choosing the method that overall shows the most robust effect sizes, and shedding light on where more research is needed. Meta-analyses that synthesize datasets across multiple topics are also possible. Finally, the goal of MetaLab datasets is that they remain as living, community-augmented meta-analyses to which researchers add new data. This allows for a cumulative approach to appraising the evidence base in a given field of research.

ACKNOWLEDGMENTS

Many thanks to the MetaLab leadership team (Alejandrina Cristia, Michael C. Frank, and Molly Lewis) for their work on MetaLab. Thank you to Page Piccinini and the MetaLab Team for their contributions to MetaLab teaching materials, which were invaluable sources for writing this paper. Thank you also to Riccardo Fusaroli and two anonymous reviewers for their helpful feedback on an earlier version of the paper. This work was supported by grants of the Institute for AI and Beyond, JSPS Grant-in-aid for Specially Promoted Research (20H05617), JSPS Grant-in-aid for Transformative Research Areas (20H05919) and a MEXT WPI research startup fund to ST, and a Metascience 2019 RFP fund by the Fetzer Franklin Fund of the John E. Fetzer Memorial Trust.

Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors declare no conflicts of interest with regard to the funding source for this study.

DATA AVAILABILITY STATEMENT

No new data were generated in support of this research. The data that were used for illustrative purposes are openly available on Open Science Framework (OSF) at https://osf.io/n9tav.

ORCID

Loretta Gasparini https://orcid.org/0000-0002-1561-5572
Sho Tsuji https://orcid.org/0000-0001-9580-4500
Christina Bergmann https://orcid.org/0000-0003-2656-9070

REFERENCES

Afshartous, D., & Preston, R. A. (2011). Key results of interaction models with centering. *Journal of Statistics Education*, *19*(3), 1. https://doi.org/10.1080/10691898.2011.11889620

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1007/978-1-4612-1694-0_16

Auguie, B. (2017). gridExtra: Miscellaneous functions for "Grid" graphics. (R package Version 2.3) [Computer software]. https://CRAN.R-project.org/package=gridExtra

Bergmann, C., Tsuji, S., & Cristia, A. (2017). Top-down versus bottom-up theories of phonological acquisition: A big data approach. In *Interspeech 2017* (pp. 2103–2107). https://doi.org/10.21437/Interspeech.2017-1443

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*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

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

Borenstein, M. (Ed.). (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Burgard, T., Bošnjak, M., & Studtrucker, R. (2021). Community-Augmented Meta-Analyses (CAMAs) in psychology: Potentials and current systems. *Zeitschrift für Psychologie*, *229*(1), 15–23. https://doi.org/10.1027/2151-2604/a000431

Carbajal, M. J., Peperkamp, S., & Tsuji, S. (2021). A meta-analysis of infants' word-form recognition. *Infancy*, *26*(3), 369–387. https://doi.org/10.1111/infa.12391

Center for Open Science. (n.d.). *Center for open science*. Center for Open Science. Retrieved August 27, 2021, from https://www.cos.io/

Center for Open Science. (2021). *Open science framework home*. Open Science Framework. https://osf.io/

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Cristia, A., Tsuji, S., & Bergmann, C. (2021). A meta-analytic approach to evaluating the explanatory adequacy of theories. [Manuscript accepted for publication]. *Meta-Psychology*. https://doi.org/10.31219/osf.io/83kg2

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536. https://doi.org/10.1037/dev0000083

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning Reviews*, *5*(1), 1–13. https://doi.org/10.1155/2012/462531

Frank, M. C., Alcock, K. J., Arias-Trejo, N., Aschersleben, G., Baldwin, D., Barbu, S., Bergelson, E., Bergmann, C., Black, A. K., Blything, R., Bohland, M. P., Bolitho, P., Borovsky, A., Brady, S. M., Braun, B., Brown, A., Byers-Heinlein, K., Campbell, L. E., Cashon, C., … Soderstrom, M. (2020). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 1–29. https://doi.org/10.1177/2515245919900809

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Irr: Various coefficients of interrater reliability and agreement*. (R package Version 0.84.1). https://CRAN.R-project.org/package=irr

Gasparini, L., Langus, A., Tsuji, S., & Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis. *Cognition*, *213*, 104757. https://doi.org/10.1016/j.cognition.2021.104757

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Griffin, J. W. (2020). MetapoweR: An R package for computing meta-analytic statistical power. (R package version 0.2.1) [Computer software]. http://CRAN.R-project.org/package=metapower

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, *11*(2), 181–217. https://doi.org/10.1002/jrsm.1378

Guyatt, G. H., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., & de Beer, H. (2011). GRADE guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, *64*(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026

Harzing, A. W. (2007). *Publish or perish*. https://harzing.com/resources/publish-or-perish

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.2307/1164588

Iverson, E., Bergmann, C., El-Shawa, S., Frank, M. C., & Tsuji, S. (2021). *MetalabR*. https://github.com/langcog/metalabr

Lajeunesse, M. J. (2016). Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution*, *7*(3), 323–330. https://doi.org/10.1111/2041-210X.12472

Laurinavichyute, A., & Vasishth, S. (2021). The (ir)reproducibility of published analyses: A case study of 57 JML articles published between 2019 and 2021 [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/hf297

Lewis, M., 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]. PsyArXiv. https://doi.org/10.31234/osf.io/htsjm

Lewis, M., Mathur, M. B., VanderWeele, T. J., & Frank, M. C. (2020). The puzzling relationship between multi-lab replications and meta-analyses of the published literature [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/pbrdk

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications, Inc.

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119. https://doi.org/10.1111/rssc.12440

Mathur, M. B., & VanderWeele, T. J. (2021). Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Research Synthesis Methods*, *12*(2), 176–191. https://doi.org/10.1002/jrsm.1464

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178. https://doi.org/10.1016/0010-0277(88)90035-2

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. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Moreau, D., & Gamble, B. (2020). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*. https://doi.org/10.1037/met0000351

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 17–29. https://doi.org/10.1348/000711000159150

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. https://doi.org/10.1037/1082-989X.7.1.105

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. (R package version 1.1-2) [Computer software]. https://CRAN.R-project.org/package=RColorBrewer

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Open Science Collaboration. (2017). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. Wiley. https://doi.org/10.1002/9781119095910.ch1

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, *15*(4), 1026–1041. https://doi.org/10.1177/1745691620906416

PRISMA. (2021). *PRISMA flow diagram*. PRISMA: Transparent Reporting of Systematic Reviews and Meta-Analyses. http://www.prisma-statement.org/PRISMAStatement/FlowDiagram

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, *22*(1), e12704. https://doi.org/10.1111/desc.12704

R Core Team. (2020). *R: A language and environment for statistical computing (3.6.3)*. [Computer software]. R Foundation for Statistical Computing. https://www.Rproject.org/

Rohatgi, A. (2020). *DigitizeR*. https://github.com/ankitrohatgi/digitizeR

Rohatgi, A. (2021). WebPlotDigitizer. (Version 4.5) [Computer software]. https://apps.automeris.io/wpd/

RStudio Team. (2020). RStudio: Integrated development environment for R (1.3.959). [Computer software]. RStudio, PBC. http://www.rstudio.com/

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., McAleenan, A., Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002. https://doi.org/10.1136/bmj.d4002

STHDA. (n.d.). *Ggplot2 box plot: Quick start guide—R software and data visualization*. *Statistical Tools for High Throughput Data Analysis*. http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization

The value of evidence synthesis. (2021). *Nature Human Behaviour*, *5*(5), 539. https://doi.org/10.1038/s41562-021-01131-7

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665. https://doi.org/10.1177/1745691614552498

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191. https://doi.org/10.1002/dev.21179

Tsuji, S., & Cristia, A. (2017). Which acoustic and phonological factors shape infants' vowel discrimination? Exploiting natural variation in InPhonDB. In *Proc. Interspeech 2017* (pp. 2108–2112). https://doi.org/10.21437/Interspeech.2017-1468

Tsuji, S., Cristia, A., Frank, M. C., & Bergmann, C. (2020). Addressing publication bias in meta-analysis. *Zeitschrift für Psychologie*, *228*(1), 50–61. https://doi.org/10.1027/2151-2604/a000393

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247. https://doi.org/10.3102/1076998609346961

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. http://www.stats.ox.ac.uk/pub/MASS4

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Von Holzen, K., & Bergmann, C. (2021). The development of infants' responses to mispronunciations: A meta-analysis. *Developmental Psychology*, *57*(1), 1–18. https://doi.org/10.1037/dev0001141

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wysocki, A., Lawson, K. M., & Rhemtulla, M. (2020). Statistical control requires causal justification [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/j9vw4

How to cite this article: Gasparini, L., Tsuji, S., & Bergmann, C. (2022). Ten easy steps to conducting transparent, reproducible meta-analyses for infant researchers. *Infancy*, *27*(4), 736–764. https://doi.org/10.1111/infa.12470