Meta-Psychology, 2022, vol 6, MP.2020.2741
https://doi.org/10.15626/MP.2020.2741
Article type: Tutorial
Published under the CC-BY4.0 license

Open data: Not Applicable
Open materials: Not Applicable
Open and reproducible analysis: Not Applicable
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Erin Buchanan
Reviewed by: Dorothy Bishop, Eiko Fried
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/2F73H

A meta-analytic approach to evaluating the explanatory adequacy of theories

Alejandrina Cristia
Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, France

Sho Tsuji
International Research Center for Neurointelligence, Institute for Advanced Studies, The University of Tokyo, Japan

Christina Bergmann
Language Development Department, Max Planck Institute for Psycholinguistics, The Netherlands

Abstract

How can data be used to check theories' explanatory adequacy? The two traditional and most widespread approaches use single studies and non-systematic narrative reviews to evaluate theories' explanatory adequacy; more recently, large-scale replications entered the picture. We argue here that none of these approaches fits in with cumulative science tenets. We propose instead Community-Augmented Meta-Analyses (CAMAs), which, like meta-analyses and systematic reviews, are built using all available data; like meta-analyses but not systematic reviews, can rely on sound statistical practices to model methodological effects; and like no other approach, are broad-scoped, cumulative and open. We explain how CAMAs entail a conceptual shift from meta-analyses and systematic reviews, a shift that is useful when evaluating theories' explanatory adequacy. We then provide step-by-step recommendations for how to implement this approach – and what it means when one cannot. This leads us to conclude that CAMAs highlight areas of uncertainty better than alternative approaches that bring data to bear on theory evaluation, and can trigger a much needed shift towards a cumulative mindset with respect to both theory and data, leading us to do and view experiments and narrative reviews differently.

Keywords: meta-analysis, variability, replication, sample size, effect size, quantitative, open science, cumulative science, theory adjudication, explanatory adequacy

Introduction

As cognitive scientists and psychologists, we strive for generality, trying to see beyond individual data points and experiments. Theories are key in this process. Ranging from broad frameworks to implemented computational models, theories are the tools we use to capture observed patterns, and to generate new predictions. Given this crucial role, theories need to be evaluated, updated, and when there are competing accounts, compared. In this context, an important question arises: How can we best evaluate theories against empirical evidence, particularly in the age of the replicability crisis (e.g., Vazire, 2018)? In this paper, we argue that usual strategies are at odds with cumulative science (defined as the endeavor to optimally integrate findings into the web of knowledge); and we propose a novel approach, based on open *community-augmented meta-analyses* (CAMAs; Tsuji et al., 2014). We first discuss the ways in which this approach more closely fits the desiderata from cumulative science, which recommends an integrative approach to empirical studies. We then provide step-by-step instructions how, in the future, we can work towards letting the evidence decide: Rather than checking a theory's explanatory adequacy via individual studies (which cannot by themselves cover the whole potential scope of a theory) or narrative reviews (where result integration is verbal), we propose a shift in mindset supported by meta-analytic tools.

Theories and cumulative science

The psychological sciences saw a sea change as reports of relatively low levels of replication bubbled to the surface (e.g., Klein et al., 2014; Open Science Collaboration, 2015). A first reaction was to blame our data collection and reporting practices: A great deal of writing has been done to quantify questionable research practices (John et al., 2012), estimating their causal impact on replication (Ulrich and Miller, 2020), and evaluating alternative research approaches (Scheel et al., 2021). More recently, attention turned to theory, with the realization that lack of replication is at least partially due to what we may call questionable theoretical practices. This young body of writing is already too extensive to be reviewed here (see e.g., Fried, 2020, and replies in the same issue), but for our purposes, the most important insights include a definition of what theory is, and what the steps of theory development are. We follow Robinaugh et al. (2021) in defining theories as models of the world, meaning that they represent in a simplified abstract manner a portion of the complexity of the world. Several researchers are in agreement about the fact that psychological theories as well as those found in many areas of cognitive science (but not all, e.g., aspects of decision-making, Palminteri et al., 2017) are purely verbal or narrative, tending to also be underspecified and ambiguous. Current recommendations are thus to strive for further precision, leading Borsboom et al. (2021) to propose that the first three phases of theory development involve 1. identifying a domain, 2. constructing a proto-theory, and 3. formalizing the theory (note that alternative proposals for steps have been laid out, for instance in Robinaugh et al., 2021; divergences on this are immaterial to the claims and proposals in the present article). Identifying a domain involves specifying the boundary of application of the theory, including the definition of its scope (i.e., when the theory applies or not). Constructing a proto-theory involves specifying what the "parts" of the theory are, as well as what their "relationships" are. In the formalization phase, the relationship between the parts comes to be defined precisely in mathematical notation. The next phase involves a check on the explanatory adequacy of the theory, a step that involves comparing the theory-implied data against empirical observations, which typically requires auxiliary hypotheses.

The present paper is focused on the phase where explanatory adequacy is evaluated. This phase has already been a focus of attention, with for instance some work explaining that this is not identical to simply fitting a statistical model to data (see Fried, 2020, pp. 274 and ss.) and arguing that instead this step will involve relating statistical modeling to data generated from formalized theories (see saliently Robinaugh et al., 2021, section 4.2). Our proposal is conceptually independent from these recommendations, as they do not specify which data bears on a theory's evaluation. We here argue that this operation should integrate all relevant and accessible information, rather than partial or select information. We detail our arguments in the next section.

How data are currently used in theory evaluation

In this section, we review two commonly employed approaches to checking the explanatory adequacy of one or more theories, against which we compare our own proposal. We assume that prior to checking explanatory adequacy, the scope of the theory has been defined, and the theory itself has been clarified and ideally formalized (see Guest and Martin, 2021; Robinaugh et al., 2021 for further information on these steps). For our purposes, what is important is that one has clarified the factors in the theory and how they can potentially be measured.

To explain our proposal, we will use a running example of how infants' learning of the sounds and words of their native language may be linked. The preceding work then, will have defined what it means by "sounds", "words", "infants", "native language", and "learn". In our running example, we will discuss three alternative (verbal) theories: top-down (stating that infants learn words first, and then use them to learn sounds; e.g., Kuhl, 1983); bottom-up (stating that infants learn sounds first, and then use them to learn words; e.g., Feldman et al., 2013); and parallel (stating that infants learn sounds and words independently from each other).

Evaluating explanatory adequacy will involve five stages. These are listed in Table 1, and we provide information on how each stage is addressed via different methods in subsequent sections. Stage I is *scope determination*, deciding which studies in the literature bear on a given theory or group of theories that are evaluated. For instance, in our running example, the research may decide that sound discrimination studies are relevant, as are word recognition studies, whereas studies where one checks whether infants prefer different prosodic patterns are not, because they do not refer to either sounds or words.

Stage II is *design space sampling*, which refers to the types of procedures, stimuli, populations, etc. that are relevant for that theory. In our running example, the researcher may decide that, given their definition of learning as changing one's behavior, behavioral studies are relevant, whereas neuroimaging studies are not. At this point, the corpus of data to be considered has been defined: it is all the studies where the "parts" of the theory are invoked, and where the "relationships" between the parts can be studied.

Stage III involves checking previous literature in terms of the *quality of the data*. There are several individual steps in this stage, which will be detailed below. The majority of checks are borrowed directly from the meta-analytic literature, and they may be involved in assessing the quality of individual studies (checking, for instance, for evidence of selective reporting of results), as well as collections of studies (checking for publication bias). In our running example, the researcher may check the quality of each body of literature they are relying on (sound discrimination studies, word recognition studies). Other checks are more specific to the evaluation of explanatory adequacy, and they involve checking whether the whole scope of the theory is already represented in the literature, or whether there are gaps in the design space that could eventually reveal inappropriate generalization. In our running example, large age gaps and differences in the age sampling for sound discrimination, which has been assessed from birth, and word recognition studies, which is not tested before the seventh month, could prevent appropriate modeling of acquisition order in the next stages.

Stage IV is *quantitatively controlling for study differences*, which may be integrated with the fifth stage discussed next, but conceptually it is closer to the third stage discussed just above. At this stage, we ask ourselves how to conceptually combine studies, given that the body of literature considered will often not be a string of strict replications or systematic variations of single factors. Some of the questions that arise require us to consider what to do with studies that vary in precision (or sample size) and/or that vary in methods (albeit within conceptual and methodological scope, given decisions made during the first and second stage). In our running example, this will entail considering what to do with studies from the 1980s and 1990s, which often had sample sizes of 6-8 infants (e.g., Kuhl, 1983), versus the 2010s, which have seen some studies with over 100 infants (Newman et al., 2016).

Stage V is *result integration*, where we try to draw a comprehensive picture based on the assembled corpus of data. When doing so, we may need to control for study differences (for instance, if several different procedures should fall within the scope of a theory, and if these lead to different results, what we are to conclude). At a minimum, this will require statistical modeling of the body of data, in which case the tools can be again borrowed from the meta-analytic literature. When theories are sufficiently specified, they may constitute theories of processes, in which case assessing results at the level of studies may be insufficient or inappropriate. In this case, the researcher will need additional modeling steps, for instance using the body of previous literature in a rawer form. Tools at this stage include Individual Participant Data (IPD) meta-analyses (Riley et al., 2020; Verhage et al., 2020); mega-analyses (Sung et al., 2014); and hybrid meta- and mega-analyses or pseudo-IPD meta-analyses (Koile and Cristia, 2021; Papadimitropoulou et al., 2019). All points discussed here apply to these different formats of quantitatively aggregating evidence, but for simplicity we limit our considerations to summarizing group-level data. In our running example, this would be when we check for evidence that the public body of literature is consistent with top-down, bottom-up, or parallel theories of early language acquisition.

Individual studies

Probably the most common way to evaluate theories' explanatory adequacy is by means of individual studies, i.e., a single experiment or manipulation (so not a paper or a series of experiments). Typically, specific predictions are empirically tested (either with human participants or computational models), and the result-

Table 1

Stages when evaluating a theory's explanatory adequacy using a single study (including large-scale replications), narrative (non-systematic) review, meta-analyses, and CAMA approaches. N/A = does not represent the whole body of literature.*

Stage	Single study	Narrative review	Meta-analyses	CAMA
I. Scope determination	N/A*	subjective, static	static	dynamic
II. Design space sampling	one point	subjective	comprehensive, narrow	comprehensive, broad
III. Checks for literature quality	N/A*	subjective	bias at study/literature-level, power analysis	bias at study/literature-level, addition of file-drawer studies
IV. Quantitatively controlling for study differences	impossible	impossible	moderator analysis, weighting	moderator analysis, weighting
V. Result integration	irrelevant	narrative; vote counting	meta-regression	replicable, reproducible, extendable meta-regression

ing data are taken to support only one of the competing accounts. For instance, in our running example, a prominent individual study often invoked as supporting the bottom-up proposal is Werker and Tees (1984), who documented a decline in the discrimination of non-native sounds between 6 and 12 months of age, before children built a vocabulary. Or so people thought at the time: Tincoff and Jusczyk (1999) found that children *did* know some words by 6 months, which put the top-down and parallel theories back in the race. But we argue that an individual study cannot be used to thoroughly check a theory's explanatory adequacy by itself, for at least the following two reasons.

First, each study is very specific: it employs one experimental setting, including stimuli, implementation, and sample, and results may not generalize to other settings that vary along one or more dimensions (Brown et al., 2014). When theorizing, we disregard the specificity of studies unless there is some other study that proves that a given setting mattered. We may then revise the theory to now predict this difference (which gives enormous weight to that result); or we may argue against the validity of that result to avoid changing our theory. This exception aside, most of the time absence of evidence of a methodological or population-specific effect is implicitly taken as evidence of absence: Each theory is as general as it can be given the extant evidence and, in return, each empirical result is taken to be as generalizable as possible barring counterevidence. We agree with Yarkoni (2020) about the fact that this is not sound theoretical evaluation practice.

Second, single studies are always a noisy window into reality. The best case scenario is that a predictable proportion of results are misleading because of our inferential tools, which allow false positives and negatives to seep into the literature. Even in this idealized case, it is impossible to determine whether a single result accurately reflects reality as there are no mechanisms to detect false positives or negatives at the study-level. To draw from our running example, data in a meta-analysis for infant vowel discrimination (Tsuji and Cristia, 2014) shows that individual studies yield a wide array of results: Across different studies, infants discriminate vowels well, barely, or not at all. However, the situation is even more complex in a realistic scenario, because it is not the case that the literature accurately reflects all findings. Indeed, extant literature (and any single study in it) may be misleading because of questionable research practices, which are eminently difficult to eradicate (Scheel et al., 2021), and because of publication bias skewed towards significant results and thus potentially over-representing false positives (Ferguson and Heene, 2012).

A special case of single studies: Large-scale replications. Recent years have seen the rise of cross-laboratory replications, which address several weaknesses we highlighted in the context of individual studies (a good set of proposals in this direction is found in Uhlmann et al., 2019). In particular, initiatives like "Many Labs" (e.g., Klein et al., 2018; Open Science Collaboration, 2015) could address both the over-specificity and the noisiness of single studies. When many labs collect data on a given phenomenon using largely the same experimental procedure they are varying experimenter identity and increasing sample diversity, which already contributes to a greater trust in the likelihood of the study generalizing to a new sample collected by a new experimenter. Their larger sample sizes also reduce the chance of observing false negatives through their greater precision. Such studies are typically also more trustworthy because analyses are usually pre-registered, and data are open, allowing correction of any analytic judgment error that may have occurred. However, these collaboration efforts have not yet gone so far as to vary methodological parameters systematically (but see Baribault et al., 2018; ManyBabies Consortium, 2020). As a result, they still provide a single datum localized to one specific region in methodological space, and thus they cannot speak to broad generalizability (see also Machery, 2020).[1]

Narrative reviews. Narrative reviews seem to provide a framework to weave together multiple studies. We talk here about non-systematic qualitative reviews, which are the prevalent form of evidence integration, often as part of the introduction and/or discussion of an experimental paper, or in invited submissions. As a result, such evaluations of the empirical evidence are often not peer-reviewed independently. Moreover, narrative reviews authored by prominent researchers come with an implied stamp of approval and are hard to contest without also appearing to attack the author – which makes the absence of appropriate peer review all the more problematic.

The first major shortcoming of narrative reviews is the fact that data selection is not done in an overt and transparent way, with no obligation to objectively check for quality and bias. In fact, despite the author's best intentions, the procedure whereby a narrative review is put together is fraught with occasions for biases to seep in, including data and outcome selection (for a self-reflective account of how this may happen, see Bishop, 2020). A documented example of this comes from a recent study of reviews on a potential link between depression and nutrition: Thomas-Odenthal et al. (2020) found strong conclusions and recommendations were eight times more common in narrative reviews as compared to meta-analyses, despite the fact that narrative reviews relied on fewer studies than meta-analyses. It may be interesting to replicate such a study focusing on a more theoretical topic in psychology.

The second shortcoming of narrative reviews is that single-study interpretation and narrativization can iron out discrepancies. For instance, going back to our running example, imagine that we find a study where infants' sound discrimination correlates with their word recognition abilities, and two studies where the correlation between the two is zero (this is based on observed patterns: Cristia et al., 2014; Wang et al., 2021). Depending on how they feel about the parallel theory, the researcher interpreting these data may argue that the latter two studies failed to find an effect because they were poorly designed or underpowered (so one piece of data supports the bottom-up account, and the other two are ignored); or they may argue that the sound discrimination study was poorly designed, loading on lexical skills, and thus this is a spurious correlation (allowing the body of results to be consistent with the parallel theory).

This is because narrative reviews lack a framework for quantitative evaluation and comparison, and thus inherit some of the issues with single studies. Sometimes, authors of narrative reviews do attempt to take into account a body of evidence with heterogeneous results – but this is hard to do in narrative terms: Authors may produce a table summarizing the studies, with a column tagging with + or - (or even 0) studies depending on whether they support a conclusion or not. This entails making a decision of what constitutes a "+" – is it a significant result, and does the direction of the effect matter? Is it a result that is numerically in the "right" direction? What is the threshold for deciding that the evidence aligns one way or another? This method is even more impractical in the case of theoretically relevant and/or methodological moderators that are suspected of having a major effect. Verbally postulating them based on diverging outcomes is not good scientific practice because it amounts to saying there is a "significant" difference without testing for it.

The final reason why narrative reviews are the most pernicious is that there is no procedure for deciding that there is *enough* evidence. Often, a single study will be considered as enough, again reflecting the "single study is decisive" assumption.

Meta-analyses. The criticisms we leveraged against single studies have motivated a push towards systematic reviews and meta-analyses in many fields, including psychology. The detailed procedures that have been laid down to guide systematic reviews and meta-analyses (e.g., PRISMA, Moher et al., 2009; Page et al., 2021; Shamseer et al., 2015) can help us counter our selection biases, overtly report quality judgments, and use objective and quantitative methods for study weighing and moderator tests. Moreover, a range of tools can be used to deal with heterogeneous data, and to check for bias in the field as a whole (e.g., Egger et al., 1997). Of course, meta-analyses are not perfect (Ioannidis, 2016), and recent investigations into the transparency and reproducibility of meta-analyses

[1] We don't discuss our running example here because there have not been any large-scale efforts to replicate sound discrimination and/or word recognition yet (but see ManyBabies Consortium, 2021).

revealed considerable issues (Maassen et al., 2020; Polanin et al., 2020). This makes sense: no tool can force its handler to use it wisely.

Meta-analyses are often done to check whether a statement is true or false – e.g., to what extent a certain treatment can reduce depression (e.g., Cuijpers et al., 2013). Considerations of moderating factors are less common (although certainly not to be ignored, see Riley et al., 2020 for the importance of integrating patient characteristics in individual participant data meta-analyses). This mindframe is appropriate for a simple hypothesis-testing, dichotomous reading of what the evidence has to tell us. As a result, heterogeneity is often seen as a threat to interpretation validity (although etiology is complex, e.g., Engels et al., 2000), meaning that some researchers will be tempted to keep the scope of their meta-analysis narrow (e.g., Li et al., 2015).

In the context of checking explanatory adequacy, such traditional meta-analyses have clear advantages over the alternative two approaches, including systematic inclusion of previous literature and overt modeling of study differences. Our running example was chosen because, in fact, there are meta-analyses for both sound discrimination (Tsuji and Cristia, 2014) and the recognition of word forms (Bergmann and Cristia, 2016), which thus provide information on the timeline of acquisition of these two levels considering all previous evidence, and statistically accounting for, e.g., methodological factors thought to be irrelevant to the theory being tested (although they account for significant variance in effect sizes; cf. Bergmann et al., 2018). Meta-analyses are, however, limited in ways that will become clear in the next section, where we explain our proposed approach.

Our proposal: CAMAs

We have proposed community-augmented meta-analyses (henceforth CAMAs; Tsuji et al., 2014) as a way to further improve on the already powerful meta-analytic approach in two key ways. First, in CAMAs the meta-analytic procedures for screening, inclusion, qualitative, and quantitative analyses, as well as the resulting data and scripts, are public and open, allowing community members to detect and correct any problems at a relatively low cost. Second, community members can rescue meta-analyses from post-publication deterioration by adding data points which emerged after the meta-analysis was originally carried out; in fact, we have seen that CAMAs provide a natural home for unpublished studies, which helps counter publication bias. Users can also add new variables of interest that the original meta-analyst might not have been aware of or interested in. As a result of these two features (openness and dynamicity), the stage is set for labor to be distributed and decision-making democratized. Extant CAMAs also suggest additional benefits. Indeed, communities are created *around* the resource for instance to profit from those data during experiment planning, leading to agreements for standardized formats to be used when extending extant CAMAs or creating new ones. This facilitates the re-use of analysis scripts and enables meta-meta-analyses.[2] Additionally, the dynamic nature of a CAMA supports the constant integration and evaluation of new evidence, naturally hijacking binary readings. We have an interesting anecdote on this which bears on our running example: When the meta-analysis on vowel discrimination was published, it was taken to support the conclusion that vowels became attuned to the native language by about 6 months, based on discrimination trajectories that were different for native than non-native contrasts, with significant increases for native ones and non-significant decreases for non-native contrasts. Since publication, the meta-analysis has become a CAMA hosted within the MetaLab platform (Bergmann et al., 2018), and last time we checked neither the native or non-native trends were significant.

Although CAMAs share many features with meta-analyses based on systematic reviews, and we can thus build on insights and methods developed (largely) in the medical sciences, their application in the context of cognitive sciences and for theory evaluation specifically does entail an important mind shift. We noted above a preference for meta-analyses to be based on a narrow scope, with heterogeneity interpreted as a validity threat. In contrast, theories in cognitive science that aim for generality will need to adopt a broader scope, which may make it burdensome for the meta-analyst (as it entails inputting more data). The unique features of a CAMA, however, help with this: Democratization of the data entry process allows other researchers to add more data points.

A step-by-step manual for using CAMAs to check theories' explanatory adequacy

In this section, we provide 10 steps you can take to check theories against extant data in a cumulative sci-

[2] See metalab.stanford.edu and PsychOpen CAMA in leibniz-psychology.org/en/services/ for implemented CAMAs. We also would like to point to Living Systematic Reviews (Elliott et al., 2014), which, as far as we can judge, are conceptually equivalent to CAMAs and were developed in parallel. In what follows, we continue using the name CAMA as this is the name under which we had proposed this idea, which has been picked up by others in the cognitive sciences (Burgard et al., 2021; IJzerman et al., 2021).

entific framework, with an extra step that is based on fostering educational synergies in research training (see Figure 1).

Step 0: Consider educational opportunities. By trying to use CAMAs to bring data to bear during theory evaluation, you will learn a great deal not only about meta-analyses, but also about how challenging it is to evaluate a theory against data in the age of cumulative science when you have not been trained for it. Considering educational opportunities means that you will make this easier for future generations of researchers, and if you think about it in advance, it may also lighten your load. Although it would be ideal to integrate early career researchers in any and all steps, the steps leading to the highest synergies are Steps 3, 4, 6, and 9. In a nutshell, these are steps in which either data are entered into a CAMA (Step 4; notice also that Steps 5-7 also involve adding information to an extant CAMA); or when you have realized data are incomplete and more needs to be collected (Steps 3, 6, 9). We have involved undergraduate and graduate students in data entry during workshops at international conferences and in our teaching (e.g. Tsuji et al., 2016; now integrated in Black and Bergmann, 2017). Regarding collecting additional data, we support the call for inviting early career researchers to be involved in replication (Hawkins et al., 2018), which would be useful to increase statistical power (Steps 3 and 9), except that we propose a twist: Instead of only engaging in strict replication, students could be involved in expanding the coverage of extant studies by varying methodology in ways predicted to be irrelevant by the theory being evaluated (Step 6). If you are teaching a data analysis course, consider using CAMAs to train students to attempt to consider data in the framework of theory evaluation statistically (using meta-regression), but also to critique extant data (by e.g. checking for design diversity, quality, heterogeneity, and power; Steps 8-10).

Step 1: Define the scope of data your theory is supposed to explain. Reflect on what you need to evaluate your theory (see Table 1): What is the range of study types the theory is thought to cover? Is your candidate theory large in scope ("how children learn") or narrow ("how adults in visual object tracking deploy overt attention")? Are there specific quality features that your theory predicts to be crucial (using a specific type of eye-tracker, removing inattentive participants)? This is also a good point at which to consider pre-registering your meta-analysis (Watt and Kennedy, 2017). As mentioned previously, for our running example on how learning sounds and words relate to each other in infancy, a reasonable scope would include studies on native sound discrimination as well as studies on processing of native words.

Step 2: Find (CA)MAs that fall within the scope defined in Step 1. Look for meta-analyses or CAMAs that match the scope you defined in the previous step. Today, this will probably involve combining multiple meta-analyses to fully occupy that scope, since most meta-analyses today are phenomenon-driven and insufficiently broad (see Step 4), and then turning your broad, composite meta-analysis into a CAMA. Meta-analyses can be turned into CAMAs, by:

- If applicable and possible, formatting raw data according to common standards (see Footnote 1)
- Providing a codebook for all columns in the raw data
- Sharing meta-analytic raw data (as extracted from papers or received from authors) and search protocols in an open format (e.g., .csv spreadsheets and .txt files) as well as code to compute effect sizes and perform analyses
- Devising a protocol for adding data (e.g., a form) and quality control (e.g., a dedicated curator)

As time goes by, the broad-scope CAMAs proposed in Step 4 will be more and more prevalent. Readers of the future: look for extant CAMAs in your fields before starting one. If you don't find one, proceed to the next step; if you do, skip to Step 5. In our running example, we found a meta-analysis on vowels and several related to word processing in MetaLab. This is a good start, but given the scope that was defined based on theoretical concerns in Step 1, we would conclude that we are missing a meta-analysis on native consonant and tone processing. It would also be ideal to have a meta-analysis on individual measures of sound and word processing (i.e., the correlation between the two).

Step 3: Stop before you start a new CAMA. You haven't found relevant CAMAs and you are uncertain whether CAMA (as an approach) is useful because there is only one or a few studies within the scope defined in Step 1. At this point, you should in fact directly conclude that *more research is needed*: If there are not enough data points to check for generalizability, how can we trust them (or it) to tell us general facts about psychological phenomena? Come back to this manual when there seems to be enough evidence. If you do find enough studies, then continue to the next step. In our running example, we established in Step 2 that there were a few relevant meta-analyses, but one estimating correlations in individual variation for sound and word processing was missing. Our own knowledge of the literature suggests that there are fewer than 5

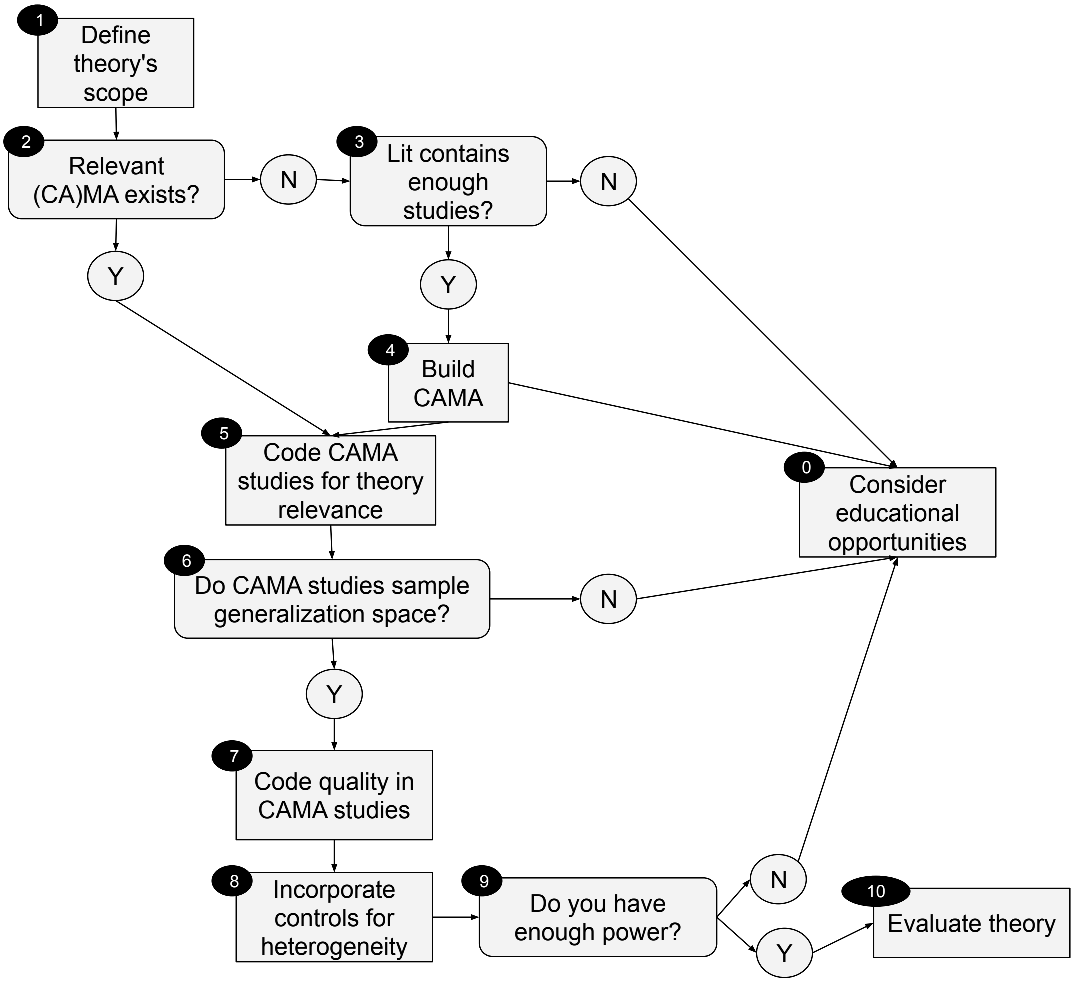
1
Define theory's scope
2
Relevant (CA)MA exists?
N
3
Lit contains enough studies?
N
Y
Y
4
Build CAMA
5
Code CAMA studies for theory relevance
0
Consider educational opportunities
6
Do CAMA studies sample generalization space?
N
Y
7
Code quality in CAMA studies
8
Incorporate controls for heterogeneity
9
Do you have enough power?
N
Y
10
Evaluate theory

Figure 1. Workflow for using CAMAs to evaluate theories. The number in the black circle refers to the Step in the manual. (CA)MA stands for (community-augmented) meta-analyses.

studies on this topic, and thus it may be too soon to attempt a meta-analysis on this topic precisely.

Step 4: Set up a broad meta-analytic scope. You've defined your scope, failed to find CAMAs that cover it completely, but believe there are enough studies to check for generalizability of the theory you are interested in, so you decide to perform a meta-analysis. Typical meta-analyses are built to evaluate whether there is sufficient evidence for a specific phenomenon, and thus data entry is limited to the scope defined by the theory. However, this means that criteria of relevance (Step 5), methodological coverage (Step 6), and quality (Step 7) are folded into one, which will make it harder to spot and recover from subjective judgments on any of these points. (Incidentally, this also limits the reusability of the data entered, and thus is in contradiction with cumulative science principles.) So think instead in CAMA terms: define your scope as broad as you can, and not any broader. In the meta-analyses we are considering for our running example, Tsuji and Cristia (2014) included all infant vowel discrimination studies (including both behavioral and neuroimaging methods, and diverse populations ranging from normative to a variety of less commonly studied infant groups); Bergmann and Cristia (2016) included all infant word segmentation studies.

Step 5: Code CAMA studies for scope. By either finding, combining, and augmenting existing (CA)MAs (Step 2) or constructing your own (Steps 3-4), you are now in possession of a body of data that probably includes studies outside of the scope defined in Step 1. Add a field to the CAMA defining relevance for your particular theory or programmatically exclude them in analysis code, for example by selecting for specific study or population characteristics. Notice that this transparency will allow future reviewers and readers to evaluate whether inclusion was subjective or principled. In our running example, the above-mentioned CAMAs on vowel and word processing were subsequently used for testing theories with a narrow scope Bergmann and Cristia (2016) and Tsuji and Cristia (2017) and a broad scope Bergmann et al. (2017). The latter was in fact an attempt to determine the relative timeline of acquisition of sounds and words. In that study, we revisited inclusion decisions: we could only find significant effects of age as predicted by the theory when we subset to studies on typically-developing monolingual infants, and which had multiple age groups in the same paper.

Step 6: Code CAMA studies for generalizability. Even after subsetting to relevant studies, the CAMA you are using may contain data collected with many methodologies. This is not a weakness. The belief that a single study can focus on a phenomenon by isolating it presumes that methodological variation goes away – which is basically an optical trick: we don't see the variation because we are focusing on one point. In contrast, broadly-defined CAMAs give us an opportunity to overtly consider that variability: Ask yourself rather, has the theory's full design space (set in Step 1) been thoroughly sampled without confounds? If so, you can use statistical tools to account for this (Step 8); if there are regions of the space that have not been sampled, or have been sampled with confounds, consider first collecting more data. In our running example, the fact that we could only retrieve the predicted age effects in a subset of data generated some concern. At present, we do not know whether this implies a true limit to generalizability of the theories, or merely a failure in statistical power, due to the fact that effects measured in infancy tend to be very small (Bergmann et al., 2018).

Step 7: Code CAMA studies for quality. As in the previous step, make sure you apply your pre-defined quality criteria from Step 1. In some research, this may mean coding whether data points come from double-blind randomized control trials as opposed to correlational research (e.g., Armijo-Olivo et al., 2015). For experimental research, you as an expert can develop field-specific criteria to code studies, ideally by crafting the definitions, and then asking a third party to apply them. An important next step is to statistically test for potential effects that confirm differences in data quality do exist. Reviewers and readers can then make an informed judgment of whether these explicit and transparent criteria were subjective or principled. Regarding our running example, we made an attempt to check whether measures of data quality defined in advance explained significant variance in the meta-analyses we were considering, and found they did not (Tsuji et al., 2020).

Step 8: Check for heterogeneity and control for orthogonal variance. In Step 1, you defined scope, design space, and quality based on the theory being evaluated. This theory may incorrectly predict homogeneity of results within this whole space. Check whether this is true using traditional meta-analytic tools, including heterogeneity checks (Huedo-Medina et al., 2006) and incorporating statistical controls for methodological (Step 6) and quality (Step 7) dimensions via weighting or as fixed or random factors, as appropriate. In our running example, we systematically control for differences in sample size by inverse variance weighting; we declare method (i.e., specific methodologies among behavioral and neuroimaging ones) as a fixed effect; and check for heterogeneity (Bergmann and Cristia, 2016; Tsuji and Cristia, 2017).

Step 9: Consider power.

At this point, you will have a CAMA covering precisely the studies within scope, sampling throughout the design space with no confounds, and taking quality into account. You are ready to integrate results using standard meta-analytic regressions, and as in such work, you should consider whether you have sufficient power (Pigott, 2020). If you find that you do not, you can estimate how much more work is needed and recommend a roadmap for future work, where you also may highlight limits on generalizability present in the extant body of literature your CAMA describes. In our running example, we found both power limitation and systematic gaps in the literature; e.g., Tsuji and Cristia (2014) found few studies on the timeline for non-native vowels, and studies since have addressed those gaps (e.g., Mazuka et al., 2014).

Step 10: Continue the work of evaluating your theory.

You have made tremendous progress in evaluating your theory in a cumulative-scientific framework – which is all the more reason to not stop now. Be extremely careful about how you interpret your meta-regression, avoiding conclusions like "the theory is (not) right because the mean effect size is (not) significant". This is once again binomial reading rearing its ugly head, now treating a meta-analysis as if it were a single study, with a focus on strict significance. Apply to meta-analyses, including CAMAs, the same lessons you learned from improved statistical practices in analyzing single experiments (see also Moreau and Gamble, 2020). In our running example, we felt that evidence at the time was most consistent with a sounds-first, than a words-first, theoretical explanation (Bergmann et al., 2017), but recognized several limitations of the evidence, including the fact that this merely indicated a difference in timelines between vowel and word processing but not a causal relationship. In any case, at this point, only one aspect of theory evaluation has occurred, and as described in the Introduction and developed further below, there are many other procedures that we can apply to not only check but also develop and improve our theories.

How this approach may change how you use single studies and narrative reviews

We believe that CAMAs are the most promising tool for transparently bringing data to bear during evaluation of theories' explanatory adequacy in the age of cumulative science. In this section, we briefly discuss the place other approaches have in the scientific process (see Figure 2).

Use CAMAs to decide not to run a new study

As CAMAs become more prevalent, it will be increasingly easy to use them to decide whether to run a new study – or not. A good example comes from our CAMA of word segmentation (Bergmann and Cristia, 2016), which documented an effect size so small that new studies have a recommended sample size of over two hundred infants, which is not currently feasible for single labs. Another example comes from a CAMA on phonotactic learning (Cristia, 2018), collecting laboratory experiments in which infants were briefly exposed to sound sequences. There were many such studies, following essentially the same method and all published as supporting the theory that prelinguistic infants can learn sound sequences after brief exposure. However, the meta-analysis revealed an effect of zero, strongly suggesting that the phenomenon was not reliable because (significant) opposite effects were sometimes observed within the same lab with nearly identical methods. This should lead at a minimum to changing the technique (habituating the child to the pattern, rather than using brief fixed exposures); and could promote an abandonment of the theory (perhaps humans can only learn sound sequences much later, after we start talking).

Use CAMA-informed single studies to efficiently sample the design space

CAMAs are useful to reveal gaps in the literature. If gathering more data along a similar line just to increase power (see Steps 0, 3, and 9), you may worry about being able to publish it. Although we do hope there is a change in attitude towards this kind of study (see also Zwaan et al., 2018), we acknowledge that such work might be most plausibly done in the context of student training, or as a first step during a PhD program (Frank and Saxe, 2012; Hawkins et al., 2018; Roettger and Baer-Henney, 2019). If collected as a student project, the sample may be too small to warrant independent publication. Nonetheless, the study would still be included in CAMAs and thus contribute to the body of evidence (see Step 7-8 for adequate integration of studies potentially varying in quality).

Use CAMA-informed studies to replicate-and-extend

Alternatively, you may be able to design your study in such a way that you both collect data that increases power on an established phenomenon, and add novel conditions, for instance to extend the coverage of methodological variables predicted to be irrelevant by the theory (see Steps 0 and 6). When writing up the results, it is then possible to emphasize the importance of the novel component (which opens the way to generating knowledge in a new direction), while calling for more work on that same topic with reference to the CAMA results. This way, an author can both signal the importance of cumulativity all the while writing a compelling article (Rabagliati et al., 2019).

Break new ground with single studies

You may have come up with a novel hypothesis for which no suitable previous data exists. Or, you may have found that extant empirical data as integrated in a CAMA contradict predictions of current theoretical accounts, and you have subjectively interpreted this contradiction, developing a new hypothesis about which factors have caused the observed discrepancy. Perhaps you decided that the method is flawed and/or the theory is false, so you would like to launch a new line of research to explore different kinds of methods and/or alternative theories. We do not want to discourage you from running this type of study. However, we hope you will remember that the role of this new study cannot be to prove or disprove a theory (see Single studies section), but to propose an idea that can then serve as a starting point for a new cumulative research endeavor.

Use narrative reviews to inform other stages of theory evaluation, adaptation, and development

We have focused here on explanatory adequacy, but as summarized in the Introduction, building solid theories takes much more than that. Some experts on this broader view of theory development recommend formalization for the precision which it brings to theoretical discussions (e.g., Robinaugh et al., 2021). Without discounting these important ideas, Guest and Martin (2021) highlight the value of considering a wide

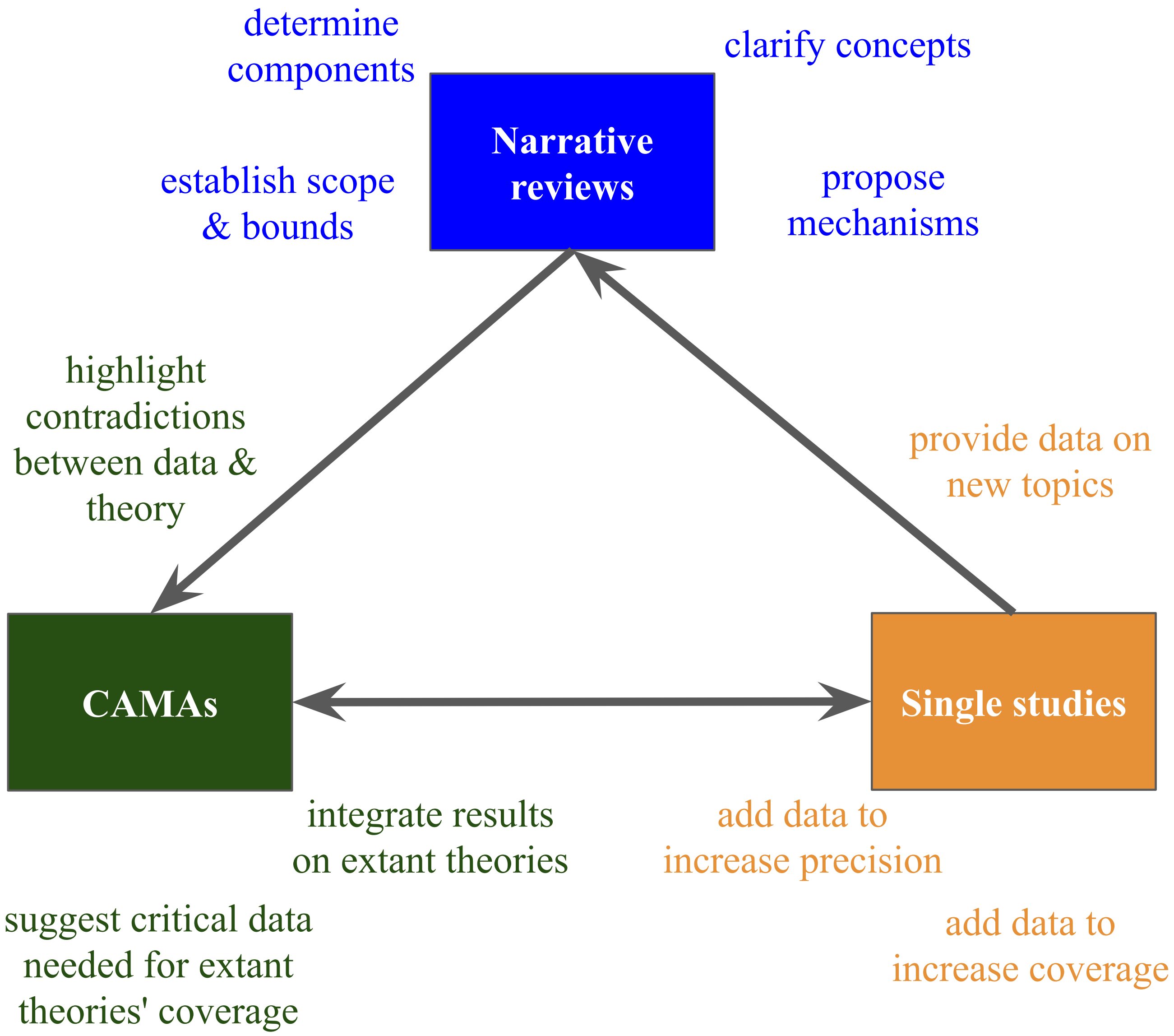
determine components
clarify concepts
Narrative reviews
establish scope & bounds
propose mechanisms
highlight contradictions between data & theory
provide data on new topics
CAMAs
Single studies
integrate results on extant theories
add data to increase precision
suggest critical data needed for extant theories' coverage
add data to increase coverage

Figure 2. Proposed key roles of CAMAs, single studies, and narrative reviews in the context of cumulative science.

range of levels of specificity when describing psychological phenomena, ranging from very specific hypotheses made in the context of one study to abstract theories in which plausible mechanisms have been specified (see saliently their Figure 2). In this context, narrative reviews still play a role as we try to clarify concepts and phenomena, and their relation to each other (crucial for theory development, e.g., Borsboom et al., 2021).

Limitations of the present paper

Before closing, we would like to highlight some shortcomings of this paper, the first being that we focused on CAMAs' role in the explanatory adequacy phase. We thus say little about other phases, and notably to the question of when one should abandon a theory altogether, which one of our reviewers cogently pointed out may be behind wasted research efforts. We believe this is an important topic that should be revisited, at which point CAMAs may be found useful in two particular ways. First, CAMAs may reveal that a theory's scope is so narrow, and/or the proportion of variance explained is so small, as to be of little use in explaining psychological phenomena in the real world. Second, having open meta-analytic repositories where data are more easily integrated into the body of literature can help provide a home for studies that would otherwise be destined to the file drawer, and thus CAMAs could help us measure wasted scientific effort.

Another limitation of the present paper is that the types of examples we have discussed are based on group-level effect sizes, typically averaged across trials and conditions, and this type of approach may be suboptimal in the quest for shedding light on cognitive processes. Haines et al. (2020) recently drew attention to this issue, and provided recommendations for data analyses. We would like to stress that systematicity, openness, and dynamicity, the three features that make CAMAs particularly powerful for testing explanatory adequacy, should carry over to this context. Of course, laying out how to engage in CAMAs using more granular data (at the trial level and below) will require additional work, which we hope will be undertaken in the future.

Conclusion

In this paper, we have considered traditional ways of bringing data to bear when evaluating theories, and concluded that none of them is perfect in the current age of cumulative science. Specifically, considering single studies in isolation (including large-scale collaboration) as well as weaving together single studies in a narrative non-systematic review both suffer from selection biases and inappropriate sampling of the space of possibilities. We have instead provided a step-by-step instruction for using meta-analyses based on rigorous systematic reviews, particularly open, community-augmented meta-analyses (CAMAs). Note that they still require the person using a meta-analysis for theory evaluation to have a clear mind about what the theory states, what its key concepts are, and what reasonable implementations of those concepts are.

Are CAMAs perfect? We suspect no, because CAMAs still rely on extant literature, and thus flaws in the literature can be carried over. Although as meta-analysts we have a few tools in our kit to deal with imperfection (see Step 8), the result of the CAMA is still bounded by the overall quality and quantity of the underlying literature, but we want to emphasize that CAMAs make the extant empirical boundaries clearer.

Being a scientist means standing on the shoulders of giants. We hope that our proposal provides guidance in how to stand firmly on these shoulders, and how others can in turn stand on ours. We look forward to a new generation of psychologists that cumulatively and systematically build on previous work, and approaches data collection and theory construction with this novel lens, making ours a sustainable discipline that ever continues to approach the truth.

Author Contact

We are grateful to Caroline Rowland, Dorothy Bishop, and Eiko Fried for invaluable feedback on an earlier version of this manuscript. All errors remain our own. Author contact:

- Alejandrina Cristia 0000-0003-2979-4556, alecristia@gmail.com
- Sho Tsuji 0000-0001-9580-4500, tsujish@gmail.com
- Christina Bergmann 0000-0003-2656-9070, chbergma@gmail.com

Conflict of Interest and Funding

The authors declare no conflict of interest related to the contents of this manuscript. The authors acknowledge 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 Nos. 659553 and 660911], Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004, ANR-14-CE30-0003, ANR-17-EURE-0017), the Fetzer-Franklin Fund, and the James S. McDonnell Foundation Understanding Human Cognition Scholar Award.

Author Contributions

AC: Conceptualization, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing; ST: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing; CB: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing

Open Science Practices

This article is theoretical and does not have accompanying data and materials, nor was it pre-registered. Thus it was not eligible for the Open Science badges. The entire editorial process, including the open reviews, is published in the online supplement.

References

Armijo-Olivo, S., da Costa, B. R., Cummings, G. G., Ha, C., Fuentes, J., Saltaji, H., & Egger, M. (2015). Pedro or cochrane to assess the quality of clinical trials? a meta-epidemiological study. *PloS one*, *10*(7).

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612.

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

Bergmann, C., Tsuji, S., & Cristia, A. (2017). Top-down versus bottom-up theories of phonological acquisition: A big data approach. *Interspeech 2017*, 2103–2107.

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.

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *73*(1), 1–19.

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. *39th Annual Meeting of the Cognitive Science Society*, 124–129.

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*, 756–766.

Brown, S. D., Furrow, D., Hill, D. F., Gable, J. C., Porter, L. P., & Jacobs, W. J. (2014). A duty to describe: Better the devil you know than the devil you don't. *Perspectives on Psychological Science*, *9*(6), 626–640.

Burgard, T., Bošnjak, M., & Studtrucker, R. (2021). Community-Augmented Meta-Analyses (CAMAs) in psychology. *Zeitschrift für Psychologie*, *229*, 15–23.

Cristia, A. (2018). Can infants learn phonology in the lab? a meta-analytic answer. *Cognition*, *170*, 312–327.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *The Canadian Journal of Psychiatry*, *58*(7), 376–385.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634.

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P., Mavergames, C., & Gruen, R. L. (2014). Living systematic reviews: An emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, *11*(2).

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in medicine*, *19*(13), 1707–1728.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438.

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*(6), 555–561.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600–604.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Haines, N., Kvam, P. D., Irving, L., Smith, C., Beauchaine, T. P., Pitt, M. A., & Turner, B. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. https://psyarxiv.com/xr7y3/download?format=pdf

Hawkins, R. X., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., Keil, M., Lampinen, A., Raposo, S., Reynolds, J., et al. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, *1*(1), 7–18.

Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, *11*(2), 193–206.

IJzerman, H., Hadi, R., Coles, N., Paris, B., Elisa, S., Fritz, W., Klein, R. A., & Ropovik, I. (2021). Social thermoregulation: A meta-analysis. https://psyarxiv.com/fc6yq/download?format=pdf

Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, *94*(3), 485–514.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524–532.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Investigating variation in replicability. *Social psychology*, *45*, 142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490.

Koile, E., & Cristia, A. (2021). Towards cumulative cognitive science: A comparison of meta-analysis, mega-analysis, and hybrid approaches. *Open Mind*, *5*, 154–173.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*(2-3), 263–285.

Li, S.-j., Jiang, H., Yang, H., Chen, W., Peng, J., Sun, M.-w., Lu, C. D., Peng, X., & Zeng, J. (2015). The dilemma of heterogeneity tests in meta-analysis: A challenge from a simulation study. *PLoS One*, *10*(5), e0127538.

Maassen, E., van Assen, M. A., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PloS one*, *15*(5).

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567.

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52.

ManyBabies Consortium. (2021). MB-AtHome: Online Infant Data Collection. https://manybabies.github.io/MB-AtHome/

Mazuka, R., Hasegawa, M., & Tsuji, S. (2014). Development of non-native vowel discrimination: Improvement without exposure. *Developmental Psychobiology*, *56*(2), 192–209.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS medicine*, *6*(7).

Moreau, D., & Gamble, B. (2020). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*. https://psycnet.apa.org/fulltext/2020-66880-001.pdf

Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *British Medical Journal*, *372*.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433.

Papadimitropoulou, K., Stijnen, T., Dekkers, O. M., & le Cessie, S. (2019). One-stage random effects meta-analysis using linear mixed models for aggregate continuous outcome data. *Research synthesis methods*, *10*(3), 360–375.

Pigott, T. D. (2020). Power of statistical tests for subgroup analysis in meta-analysis. *N. Ting, JC Cappelleri, S. Ho,(Din) D.-G. Chen (editors), Design and Analysis of Subgroups with Biopharmaceutical Applications*, 347–368.

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.

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).

Riley, R. D., Debray, T. P., Fisher, D., Hattle, M., Marlin, N., Hoogland, J., Gueyffier, F., Staessen, J. A., Wang, J., Moons, K. G., et al. (2020). Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning. *Statistics in medicine*, *39*(15), 2115–2137.

Robinaugh, D. J., Haslbeck, J. M., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743.

Roettger, T. B., & Baer-Henney, D. (2019). Toward a replication culture: Speech production research in the classroom. *Phonological Data and Analysis*, *1*, 13.

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *British Medical Journal*, *349*.

Sung, Y. J., Schwander, K., Arnett, D. K., Kardia, S. L., Rankinen, T., Bouchard, C., Boerwinkle, E., Hunt, S. C., & Rao, D. C. (2014). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. *Genetic epidemiology*, *38*(4), 369–378.

Thomas-Odenthal, F., Molero, P., van der Does, W., & Molendijk, M. (2020). Impact of review method on the conclusions of clinical reviews: A systematic review on dietary interventions in depression as a case in point. *PloS one*, *15*(9).

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological science*, *10*(2), 172–175.

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

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

Tsuji, S., & Cristia, A. (2017). Which acoustic and phonological factors shape infants' vowel discrimination? exploiting natural variation in inphondb. *INTERSPEECH*, 2108–2112.

Tsuji, S., Cristia, A., Frank, M. C., & Bergmann, C. (2020). Addressing publication bias in meta-analysis. *Zeitschrift für Psychologie*, *228*, 50–61.

Tsuji, S., Lewis, M., Bergmann, C., Frank, M., & Cristia, A. (2016). Tutorial: Meta-analytic methods for cognitive science. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Roceedings of the 38th annual conference of the cognitive science society* (pp. 33–34). Cognitive Science Society.

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific utopia iii: Crowdsourcing science. *Perspectives on Psychological Science*, *14*(5), 711–733.

Ulrich, R., & Miller, J. (2020). Meta-research: Questionable research practices may have little effect on replicability. *Elife*, *9*.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417.

Verhage, M. L., Schuengel, C., Duschinsky, R., van IJzendoorn, M. H., Fearon, R. P., Madigan, S., Roisman, G. I., Bakermans–Kranenburg, M. J., Oosterman, M., & on Attachment Transmission Synthesis, C. (2020). The Collaboration on Attachment Transmission Synthesis (CATS): A move to the level of Individual-Participant-Data meta-analysis. *Current Directions in Psychological Science*, *29*(2), 199–206.

Wang, Y., Seidl, A., & Cristia, A. (2021). Infant speech perception and cognitive skills as predictors of later vocabulary. *Infant Behavior and Development*, *62*.

Watt, C. A., & Kennedy, J. E. (2017). Options for prospective meta-analysis and introduction of registration-based prospective meta-analysis. *Frontiers in Psychology*, *7*, 2030.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.