Can infants learn phonology in the lab? A meta-analytic answer

Alejandrina Cristia

LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL Research University, France

ARTICLE INFO

ABSTRACT

Two of the key tasks facing the language-learning infant lie at the level of phonology: establishing which sounds are contrastive in the native inventory, and determining what their possible syllabic positions and permissible combinations (phonotactics) are. In 2002–2003, two theoretical proposals, one bearing on how infants can learn sounds (Maye, Werker, & Gerken, 2002) and the other on phonotactics (Chambers, Onishi, & Fisher, 2003), were put forward on the pages of *Cognition*, each supported by two laboratory experiments, wherein a group of infants was briefly exposed to a set of pseudo-words, and plausible phonological generalizations were tested subsequently. These two papers have received considerable attention from the general scientific community, and inspired a flurry of follow-up work. In the context of questions regarding the replicability of psychological science, the present work uses a meta-analytic approach to appraise extant empirical evidence for infant phonological learning in the laboratory. It is found that neither seminal finding (on learning sounds and learning phonotactics) holds up when close methodological replications are integrated, although less close methodological replications do provide some evidence in favor of the sound learning strand of work. Implications for authors and readers of this literature are drawn out. It would be desirable that additional mechanisms for phonological learning be explored, and that future infant laboratory work employ paradigms that rely on constrained and unambiguous links between experimental exposure and measured infant behavior.

1. Introduction

One of the fundamental tasks facing human infants pertains to learning the ambient language's phonology, including both determining the sound inventory and the constraints regarding the position and sequencing of those sounds. An overwhelming body of evidence suggests that attunement to the ambient language's phonology begins in infancy (Werker & Hensch, 2015), raising the question of what mechanisms may begin to operate at such an early age. Given that infants' lexicon is highly constrained, a mainstream assumption posits that infants may begin to learn their language's sound system by applying simple statistical mechanisms directly onto the spoken signal they hear.

For learning native sound categories specifically, Maye, Werker, and Gerken (2002) proposed that infants track the distribution of acoustic cues in the input, as modes or peaks in the distribution of one acoustic cue could reflect the presence of a sound category implemented in that location of acoustic space. The underlying intuition runs as follows: if a language has two sounds that contrast along an acoustic dimension, then there will be two peaks in the distribution of tokens along that

dimension (one corresponding to each sound), whereas if there is only one category in those regions of acoustic space, there will be a single peak in the frequency distribution. A sensitivity to modes in the frequency distribution would then help infants learn the sound system of their ambient language. For learning phonotactics (the regularities concerning the position and sequencing of sounds), Chambers, Onishi, and Fisher (2003) proposed that infants keep track of the frequency with which sounds occur in a given syllabic position and/or in a specific order.

These interesting proposals can be evaluated in a number of ways. One crucial way of assessing the potential explanatory value of such proposals is to carry out proofs of principle demonstrating that, at least in ideal situations, infants extract the kind of information one predicts they would. For those two theoretical proposals, this has been done using what may termed artificial grammars, languages, or phonologies: Researchers devised a simplified language to represent some feature of human language, exposed infants to some exemplars generated from that grammar, and later tested infant perception with new items that could be distinguished if the artificial phonology had been learned.[1]

---

[1] It may be relevant to point out that the artificial phonologies constructed for these studies are not only much simpler than those of natural languages, but may not have the same properties. For example, a key notion of phonology is that of minimal pair, two words that are identical except for one sound, yet mean different things; none of the artificial phonologies built for these studies contain any reference to meaning. Furthermore, there is no evidence that infants learn them using the exact same mechanisms they employ to learn natural languages' phonologies. Nonetheless, we have no evidence that they do employ the same cognitive mechanisms available for learning other types of categories and regularities outside the domain of phonology, in the auditory or any other modality. Instead, the choice of the term 'artificial phonology' reflects what the experimenter put in, rather than what the infant actually does. The actual learning may turn out to be as general as 'auditory learning' (applying equally to speech and non-speech) or as specific as 'adaptation to a syllable repeated more than 20 times' (and not applicable to any other situation).

Indeed, the original theoretical proposals in Maye et al. (2002) and Chambers et al. (2003) were supported each by two such experiments. Additionally, the same authors and others continued to explore the proposed basic mechanisms in additional laboratory-based experiments that could be considered conceptual replications of the original studies (e.g., Liu & Kager, 2014; Maye, Weiss, & Aslin, 2008; Wanrooij, Boersma, & Zuijen, 2014; Yoshida, Pons, Maye, & Werker, 2010, all followed up on Maye et al., 2002's sound category learning proposal).[2] Undoubtedly, such experiments are only the first step in laying out the explanatory power of such proposals. For instance, additional steps may include checking whether the spoken input available to infants contains the characteristics attributed to it (i.e., whether the number of modes in acoustic cue distributions corresponds to the number of sound categories in the language; Bion, Miyazawa, Kikuchi, & Mazuka, 2013); and designing computational models that mimic the procedure attributed to the learner, and assessing whether such a learner can succeed in detecting the native phonological inventory when presented with idealized or realistic input (e.g., Vallabha, McClelland, Pons, Werker, & Amano, 2007; Versteegh et al., 2015). Those two proposals have inspired a great deal of research, leading to numerous experimental extensions, and inviting both corpora studies and computer models. In view of the important theoretical implications of Maye et al. (2002) and Chambers et al. (2003), as well as their widespread effects in the study of early language acquisition, the present paper seeks to assemble extant empirical evidence on such proofs of principle, and assess how convincing this evidence is through meta-analytic tools.

1.1. Why carry out a meta-analysis?

Some readers may wonder: How can a meta-analytic approach be integrated into research on cognition? A single experiment sometimes appears more compelling than the accumulation of varied evidence. Indeed, well-designed, robust experiments can be extremely useful in shedding light on cognitive processes whose effects are fundamental, yet so slight or transitory that they are not obvious in simple observations of behavior in the 'wild'. What a meta-analytic approach can add is an estimation of how *robust and replicable* an experiment is. When we carry out a laboratory experiment, the assumption is that we are creating replicable conditions for observing the effects of a hypothesized cognitive construct. If those conditions are re-created at another point in time, or by a different group of researchers, provided that the cognitive construct is available to that second group of infant participants, we as experimentalists would predict that we will observe similar effects. Naturally, since any single experiment is a noisy observation of underlying reality, we can expect that effect sizes will vary across experiments, and can then use a cumulative approach to estimate the underlying effect size despite variations due to noise.

A meta-analytic approach can further help us assess whether variance in results across conceptual replications is systematic. For instance, it can help us measure the effects of changes in the design across different implementations of the same general conceptual goal, or test the reliable effect of a factor that has been invoked on theoretical or empirical grounds. For example, infant age is often discussed when

attempting to explain changes in performance, as older infants may be more entrenched in their native language and more resistant to the exposure (Yoshida et al., 2010), and younger infants, although more flexible, may be cognitively limited and learn more slowly (Seidl, Cristia, Onishi, & Bernard, 2009). These differences often occur within the same paper (Seidl et al., 2009), but sometimes those factors are invoked to explain differences across data sets published in different papers (Cristia, Seidl, & Gerken, 2011). If age is a variable that structures performance, then it should explain some variance in results even when age does not vary by design, and should thus be evident when comparing the size of the effects found in different experiments, including those that are unrelated to the scientists making the initial claim.

Finally, meta-analyses can uncover a further source of structure in public data, namely that emerging from conscious or unconscious biases affecting the producers of those data. The psychological sciences are seeing today a revival of concern in the robustness and replicability of our results. The problem likely emerges from the fact that the current reward scheme pushes researchers, reviewers, and editors to value more results where there is a p-value below the .05 threshold than results that are not significant (see e.g., Ioannidis, 2005; Nosek, Spies, & Motyl, 2012 for the general argument, and Open Science Collaboration, 2015 for a recent set of results in psychology). As a consequence, there is an over-representation of significant results in the literature, and quite likely an increase in the number of false positives that are present in published studies (Sterling, Rosenbaum, & Weinkam, 1995). Beyond the why's and how's, what is certain is that the community should be careful when interpreting published literature, as it may be the joint result of actual findings and biases. A meta-analytic approach can help us determine whether there is evidence of biases in reporting, through the study of the systematic patterns found in the literature.

1.2. The present work

My main goal is to inform readers about the overall empirical value of artificial phonology studies for our understanding of putative language learning mechanisms in infancy, covering distributional learning for establishing the phonological inventory on the one hand, and phonotactic learning on the other. A secondary goal, which will be evident in the discussion, is to explore implications of the results found for the empirical literature on infant laboratory learning at large.

To address the primary goal, I apply meta-analytic tools to the body of infant artificial phonology studies in order to describe it in three ways. First, I assess the overall robustness of effects found, estimated through the weighted mean effect size. Second, I explore whether certain factors are moderators of overall effect sizes, meaning that there is variance that may be systematically attributed to them, for example design characteristics that regularly lead to greater or smaller effect sizes. Finally, I investigate the possibility that there is selective reporting in this literature via inspection of funnel plots and reported p-values, techniques that will be introduced in further detail below. Although for conceptual reasons one may prefer to present the sound category learning literature before the phonotactic learning one, the methods are more complex in the former than the latter. Therefore, I will present them in the opposite order to facilitate readers' comprehension.

2. Phonotactic learning

I have followed the 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses' (PRISMA) statement (Moher, Liberati, Tetzlaff, & Altman, 2009). A file with inclusion/exclusion decisions, a flow-chart summarizing this process, the PRISMA checklist, a spreadsheet containing the meta-analytic data, and all analysis scripts are available for download from https://osf.io/9zd2a/.

---

[2] A note may be needed regarding the word "replication". Some hold that a replication is only an experiment that keeps all factors constant; in the extreme, the only way to replicate would be to go back in time and run the experiment again, as it is possible that changes in e.g., weather, society, prevalence of mild otitis, etc. could affect results via their effects on the infants tested. More commonly, an experiment constitutes a *strict* replication when the original experiment is repeated on a new sample of the same population, in the same or another lab. Finally, the term *conceptual* replication is used for cases in which only the conceptually key factors are kept constant, but methods are allowed to vary. Studies included in the present meta-analyses fall in the latter category, as stimuli, infant age, and sometimes preference method can change between the original study and the follow-ups, but they all share the same conceptually key events: exposure followed by some test in which preference could only be explained if infants had picked up on an underlying contrast or regularity.

2.1. Methods

2.1.1. Paper identification

I use 'paper' as a blanket term including articles, conference proceedings and abstracts, book chapters, and unpublished reports. The paper pool was composed by combining a list of papers known to the author with systematic searches on scholar.google.com and pubmed. Additionally, I have included unpublished data that I or others have collected in the search. The detailed timeline and methods for study identification can be found on https://osf.io/wxbdp/.

2.1.2. Paper selection

It is important in meta-analyses to not mix 'apples and oranges'. Therefore, only near conceptual replications of the seminal study by Chambers et al. (2003) have been included. The inclusion criteria were: (a) Participants are typically-developing children, between the ages of 0 and 36 months; (b); participants experience a passive exposure in the lab, and are subsequently tested via any behavioral or non-behavioral method also in the lab; (c) The exposure is to word-like items that contain certain phonotactic restrictions, the test consists of new word-like items that follow or violate those restrictions.

There were three types of test trials, depending on the relationship between the sounds used in familiarization and test. I will call a trial 'familiar' if a sound used in the exposure phase followed at test exactly the pattern that it had been used to evidence during exposure; legal if a novel sound was used, but it was a plausible generalization from the set of sounds used in the exposure and in the appropriate sequences; and illegal if the phonotactic patterns common in the exposure items were violated.[3]

After applying these criteria, there were 16 papers that could be considered for a quantitative meta-analysis, from a variety of sources: 9 journal articles (Chambers et al., 2003; Chambers, Onishi, & Fisher, 2011; Cristia & Seidl, 2008; Gerken & Knight, 2015; Gerken & Quam, 2016; Seidl & Buckley, 2005; Seidl et al., 2009; Seidl, Onishi, & Cristia, 2014; Wang & Seidl, 2015), 2 articles in proceedings (Cristia & Peperkamp, 2012; Cristia et al., 2011), 2 theses (Chambers, 2004; Cristia, 2006), 1 chapter in a collection (Cristia, Seidl, & Francis, 2011), 2 sets of data that have not been published (one reported on in Cristia, 2015).

As noted above, an effort was made to retrieve relevant unpublished work. Whether meta-analyses should include unpublished data is a subject of controversy, given that arguments for and against it can be posited (see a clear summary in Smith & Egger, 1998). On the one hand, biases may be compounded in journal publications, as the publication process entails pressures from reviewers and editors to remove non-significant or counterintuitive results that the original author might have wished to include. On the other hand, unpublished information from conference abstracts, for instance, may be just as liable to selective reporting and inappropriate analyses, and they may further be based on lower quality or incomplete data. The latter problem is particularly salient when reports are public and not carefully peer-reviewed, and when authors have a conflict of interest. These considerations seem less relevant in the context of infant research, when the same labs provide both published and unpublished data, as is the case for the meta-analyses in the present paper. Nonetheless, given the general uncertainty, I have carried out analyses over the full dataset as well as focusing on journal-published work. For simplicity, I report only on the latter here (see online supplementary analyses https://osf.io/gdsg9/ for a meta-analysis based on all studies); both analyses yield the same patterns of results.

Three additional studies were considered but ultimately excluded. Two of them investigated effects of phonotactic learning on sensitivity to a sound contrast, such that infants come to be less sensitive to a pair of sounds when they are in complementary distribution than when they can occur in the same phonological positions (White, 2014; White, Peperkamp, Kirk, & Morgan, 2008). Thus, the two papers that use this method conceptually bear more on the effects of phonotactic learning than the acquisition of phonotactic patterns, and empirically use a testing phase that is very different from all other phonotactic studies. The third was a neuroimaging study, with a rather different presentation procedure compared to the others (Obrig et al., 2017). During pre- and post-test a set of pseudowords were presented to 6-month-olds; some pseudowords contained phonotactic patterns absent from the infants' native language, whereas others respected the infants' language phonotactics. Half of each were again presented during a training phase, which consisted of exposure to pseudo-words paired with visual objects. Because the study included both native-legal and native-illegal phonotactics, and tested infants on pseudowords that had been part of the training and others that had not, a three-way interaction (specifically, the contrast in brain responses to pre- versus post-training for the native-illegal non-trained pseudowords) could have been seen as reflecting phonotactic learning and generalization. Nonetheless, the study seemed different enough, conceptually (as phonological patterns could be extracted from the newly-acquired lexicon) and methodologically (e.g., the use of a pre- and post-test), to warrant exclusion.

2.1.3. Data entry

All papers were entered and checked a second time by the author. Some papers report on more than one experiment. In this case, there is one 'record' per experiment, because each experiment can potentially provide one effect size. All experiments have used a counterbalanced design whereby one set of infants is exposed to rule A and thus should show a preference among A and -A items; the other group hears -A and should show the opposite preference. Whenever results were reported separately for counterbalanced conditions, they were separated and otherwise they were collapsed.

A total of 36 journal-published records were thus identified. Of these, in 4 cases the effect sizes emerged from a single group of infants tested with illegal, legal, and familiar test trials, and which therefore contributed data for two types of cross-trial contrasts, illegal-familiar and illegal-legal. Repeated measures are more likely to be similar than measures drawn from independent samples, simply because the repeated measures come from the same individuals. Therefore, in the standard meta-analytic practice followed here, the effect sizes emerging from repeated measures were averaged together such that each group of infants was only counted for one effect size (i.e., a total of 2 effect sizes), yielding a final total of 34 independent, comparable, journal-published effect sizes.

Each record was coded in terms of a number of dimensions, of which the following are relevant for the analyses and presented here (see supplementary materials for further details and analyses):

- background information on the paper, such as the year of publication or completion and publication type
- average age of the infants
- the number of infants included and excluded
- the type of test trials contrasted (e.g., comparing looking times to illegal versus familiar trials; or comparing looking times to illegal versus legal)
- the mean and standard deviation of looking times for each trial type
- the correlation in individuals' looking times across the two trial types within each experiment

Additionally, I coded papers on a number of other dimensions that have been found to be relevant in previous meta-analyses of infant laboratory results (Bergmann & Cristia, 2016; Cristia, Seidl, Junge,

---

[3] I followed the original authors' judgments regarding what counted as plausible generalization sounds. Admittedly, how and when listeners generalize from a set of sounds used in the exposure to other sounds is an open topic – see e.g., Moreton and Pater (2012) for a recent review of artificial grammar learning work.

Step 1: r_{FN} imputation

| Non-missing values randomized with replacement... | ... and attributed to missing values |

r_{FN-1} r_{FN-6} → r_{FN-1}

r_{FN-2} r_{FN-3} $r_{FN-6\ imputed}$

r_{FN-3} → r_{FN-3} r_{FN-3}

...   .

r_{FN-n} r_{FN-n} r_{FN-n}

Step 2: Pooled SD $SD_P = \sqrt{(SD_F^2 + SD_N^2)/2}$

Step 3: Effect size $d = \dfrac{\overline{x}_N - \overline{x}_F}{SD_P}$

Step 4: Hedges' correction $g = d * \left(1 - \dfrac{3}{4 * N - 9}\right)$

Step 5: Standard error of the effect size $SE_g = \sqrt{\dfrac{2 * (1 - r_{FN})}{N} + \dfrac{g^2}{2 * N}}$

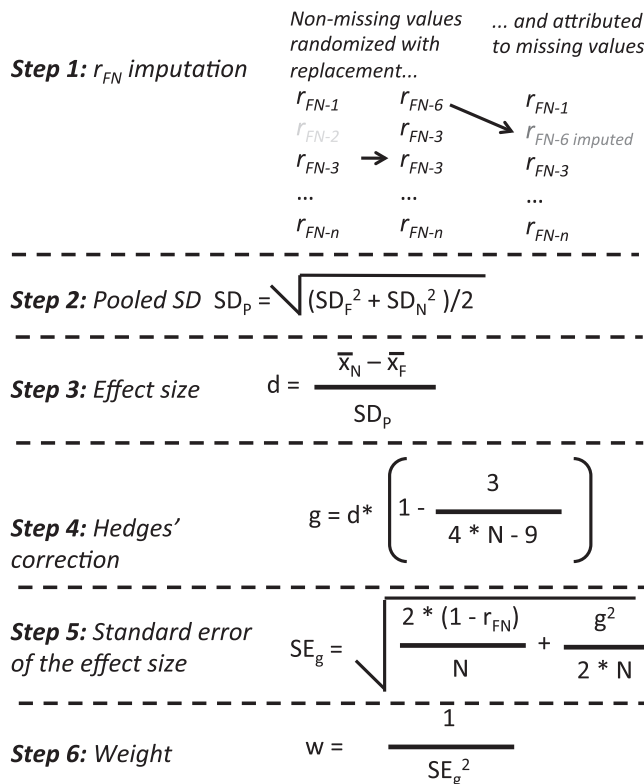**Step 6:** Weight $w = \dfrac{1}{SE_g^2}$

Fig. 1. Procedure used to impute correlations and calculate effect sizes, standard errors of the effect size, and weights. r_{FN} is the correlation in individual looking times during familiar-like and novel-like trials, with subscripts for each of the records (light gray is used to represent values that were not available); x_F is the average of looking times to more familiar stimuli (e.g. legal in a legal-illegal contrast); x_N the average of looking times to novel-like stimuli (i.e., illegal in the same contrast); $SD_{F/N}$ their respective standard deviations and SD_P the pooled SD; N the number of infants included; d effect size Cohen's d; g effect size Hedges' g; SE_g the standard error of the effect size; w weight.

Soderstrom, & Hagoort, 2014; Tsuji & Cristia, 2014), and which would be useful in a meta-meta-analysis (Lewis et al., 2015). A complete explanation for all fields can be found in https://osf.io/ge8wt/.

2.1.4. Effect size calculations

Using the coded information, an effect size, the standard error of the effect size (which is partly derived from sample size), and a weighting factor were calculated for each experiment as follows (see Fig. 1).

The **effect size** is defined in general terms as the ratio of the mean difference divided by the pooled standard deviation (SD). For the present meta-analysis, the dependent measure was looking time to two types of trials, and in every case one of the trials can objectively be described as being more novel than the other. That is, infants who are tested with legal versus illegal should find the illegal more novel; those who are tested with familiar (i.e. using sounds whose phonotactics has been evidenced in the initial exposure) versus legal (i.e., sounds that are plausible generalization targets) should find the legal trials more novel, etc. This coding scheme was kept constant across all records (cf. Cristia & Peperkamp, 2012, caption under Table 1), and thus all looking times were entered into two columns, one for more novel trials and the other for more familiar trials. As a result, the mean difference used as a numerator in the effect size is the difference between the mean looking times to the more novel trial type minus that for the less novel trial type (e.g., mean for illegal minus mean for legal), and the pooled standard deviation collapses across trial types. Finally, effect sizes were scaled using a factor that is dependent on the sample size following Hedges' recommendations (Hedges, 1981), to derive the effect size measure 'g'.

The **standard error of the effect size** depends on the sample size, the effect size, and the correlation between the two within-participants

repeated measures (i.e., higher correlations in individuals' looking times across trial types suggests a more powerful and precise effect size). Authors shared raw data allowing the calculation of 23 correlations, out of the 48 records from any venue. Following standard meta-analytic practice (e.g., Lipsey & Wilson, 2001), the remaining values were imputed.[4] Finally, the effect-size standard error is inverted to derive the **weight** of individual studies.

2.1.5. Analyses

All analyses have been carried out in R (R Core Team, 2015) and rendered in this manuscript using knitr (Xie, 2014) and xtable (Dahl, 2009). Three key questions are addressed, using mostly the metafor package (Viechtbauer, 2010). First, to establish the overall strength of the effects of laboratory phonological learning a random effects meta-analytic regression is fit (a regression where observations are experiments rather than individuals).[5] The intercept can be viewed as the weighted mean of the effect sizes, which represents here the basic phonotactic learning effect, meaning the extent to which looking time preferences can be stably induced by exposure to a phonotactic pattern.

Our second question concerns variance explained by conceptual and methodological features. A second meta-analytic random effects regression was fit, this time declaring such dimensions as moderators. The factors that (could) have been invoked to explain divergences in results, and which varied sufficiently in the sample, were:

- Age: younger infants may be less able to encode phonotactics, resulting in familiarity (rather than novelty preferences; Seidl et al., 2009), or complete failure for complex rules (Chambers et al., 2011);
- Rule complexity: the phonotactic pattern involves a (more complex) second-order dependency rather than a (simpler) first-order dependency (sound predicted from syllable position; Chambers et al., 2011);
- Number of tokens and types presented during the exposure: infants may have a harder time when provided with fewer word types/tokens during the familiarization (Cristia & Peperkamp, 2012);
- Test type: it is conceivable that it is harder to discriminate between legal and illegal trials (which are all novel to the child) rather than between familiar and illegal trials (the former using sounds already present in the familiarization).

Neither age nor number of types were distributed normally in this sample. In the case of age, no transformation seemed appropriate to capture the data distribution, which had a peak at around 10 months with a wide spread between four and 16 months. Therefore, we should be cautious when interpreting results for this factor. For number of

---

[4] This imputation can be done in a number of ways; for example, by treating all missing correlations as zero, which reduces the power of the meta-analysis (since treating it as zero is the same as treating the looking times to legal/familiar and illegal as if they were coming from different infants, rather than the same infants). Instead, I used randomization with replacement, which allows one to provide an estimation of the missing values that is informed by the knowledge from other studies, which are likely similar to those having missing values. This process, illustrated at the top of Fig. 1, replaces each of the 25 missing correlation values with one drawn at random from the 23 correlations that are available. Notice that this imputation procedure only affects, and in a slight manner, the relative weight of studies. In supplementary materials, I show that my main conclusions are not affected by following this procedure, compared to using a correlation of zero for all studies with missing values (see https://osf.io/gdsg9/, pp. 9–13).

[5] It is unclear whether a fixed effects or a random effects regression should be used here. Since the studies are all conceptual replications and follow the methods of the seminal study (Chambers et al., 2003) very closely, and all of them fit in neatly into the definition of experiments, a fixed effects model, which has the benefit of being more powerful, may be appropriate. However, I concluded that a random effects model was more informative to the readership because it theoretically allows one to consider the generalizability of findings to unobserved points, and because it does not assume that the exact same effect underlies all observations. That is, although all experiments included are extremely similar to each other, there are nonetheless differences across them, and they cannot therefore be viewed as strict, perfect replications.
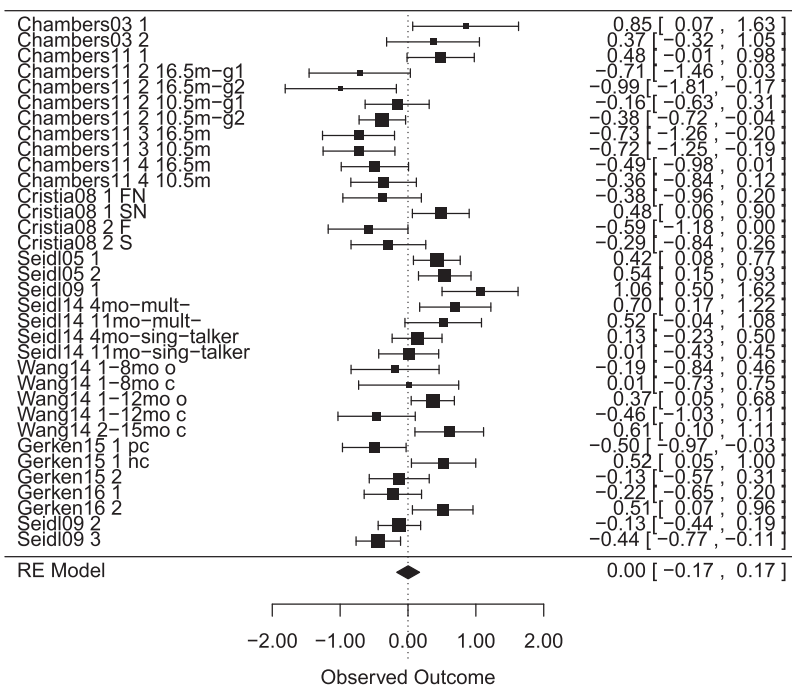
Fig. 2. Forest plot of the phonotactic learning meta-analysis. Each row represents an effect size representing infants' looking to novel-like trials minus familiar-like trials. The center of the square for that line indicates the observed effect size and its size represents the weight, whereas the whiskers delimit the confidence intervals. The diamond represents the overall effect.
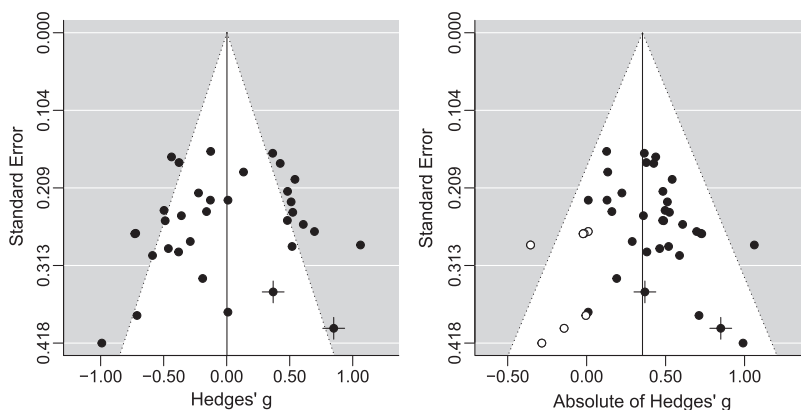

Fig. 3. Funnel plots of the phonotactic learning meta-analysis showing, for each independent sample, the standard error of the effect size as a function of the raw effect size, on the left panel, and absolute effect size, on the right panel. The white circles have been added by the trim-and-fill method in an attempt to symmetrize the distribution (no such additions were necessary for the left-hand funnel plot). The two samples with an overlaid cross are those reported on in the seminal study by Chambers et al. (2003).

types, three groupings were created, representing low (4–16 pseudo-words), medium (24–26 pseudowords), and high (48–60 pseudowords) variability. It is also important to point out that these three factors are not empirically separable in the present sample (e.g., there is no high variability + second-order + illegal-legal contrast study).

Our final research question concerns systematic patterns in experimental results that may be indicative of research practices leading to a biased literature. Such practices include deciding on excluding data and/or increasing the sample size after inspecting the results of statistical tests; performing repeated significance testing with slightly different parameters until the p-value falls below the alpha threshold; and withholding certain experiments when they are not significant or trend in an unexpected direction. The current consensus in psychological science is that these practices seriously compromise the integrity of published data (e.g., Simmons, Nelson, & Simonsohn, 2011). There is no reason to believe that the research that is being meta-analyzed here, and research on infant cognition at large, is entirely free from such

practices (as documented by Peterson, 2016). Since we only have access to public results, it becomes crucial to establish to what extent these public results are unbiased reflections of the underlying effects.

First, meta-analytic data was described with a funnel plot (see Fig. 3 below for an example). Such plots show the standard error of the effect size as a function of the effect size, with the x-axis centered in the weighted average effect size (which is our best estimate of the underlying effect size) and the y-axis plotted such that more precise observations are higher up and less precise ones are closer to the bottom of the graph. An intuitive way of thinking about the 'triangle' is that precise studies, with small standard error and thus higher up along the y-axis, tend to have many participants. If a lab runs many participants in a study, it is making such a large investment that authors are more motivated to find a way of rendering results public. In contrast, if a lab has only ran a few participants in a study, the investment has been smaller and the associated authors may be less motivated to publish the results regardless of their direction. This allows the bottom of the

Table 1
Distribution of reported p-values below alpha = .05 in the literature on infant laboratory learning of phonotactics.

(0,.01]	(.01, .02]	(.02, .03]	(.03, .04]	(.04, .05]
7	4	2	2	5

triangle (lower precision and smaller sample sizes) to be sparser than the top in selective ways, typically with more results right of the mean effect size (i.e., overestimating the effect).[6] To illustrate selective data loss, one can use a "trim and fill" method, which generates additional points from observed data as needed to symmetrize the funnel plot.

Second, I inspected the distribution of p-values. It has been proposed recently that, if authors make post hoc decisions based on inspection of the data to meet the significance criterion of alpha = .05 ("p-hacking"), then results just around this cutoff will be over-represented compared to significant results that are much lower than this cutoff (e.g., Head, Holman, Lanfear, Kahn, & Jennions, 2015; Simonsohn, Nelson, & Simmons, 2014). If the distribution of significant p-values shows many observations around p = .05, then this would be consistent with p-hacking being common in that body of literature. In contrast, if there are many observations with very low p-values, this distribution would be diagnostic of strong evidential value.

2.2. Results

Inspection of the distribution of 34 independent effect sizes reported on in journal articles (and therefore included in this meta-analysis; see Appendix A for the full list), revealed that no data point was more than 3 standard deviations away from the group mean, and thus no outliers needed to be excluded.

In answer to the first question, a random effects model declaring no moderators revealed that the intercept was not significantly different from zero: effect size $g = 0.002$, SE of the effect size g intercept = 0.086, $z = 0.025$, $p = 0.98$ (see forest plot in Fig. 2). The test for heterogeneity in effect sizes was highly significant, $Q(33) = 129.399$, $p < .001$, strongly suggesting that there was unexplained variance in effect sizes found across the different studies.

Additional models were fit to answer the second research question by declaring the conceptual moderators described in the Methods section. Overall, the moderators did not explain significant variance [$QM(6) = 6.704$, $p = 0.349$; see Appendix A.2, Table 5, for the precise estimates for each moderator].

The last set of analyses focuses on evidence for bias in the literature. The funnel plot asymmetry test was not significant ($z = -0.788$, $p = 0.431$).

As a reviewer pointed out, however, these systematically-coded effect sizes do not represent what authors, reviewers, and editors in infancy research commonly consider as the result of a looking-time study. Indeed, common practice maintains that either direction of preference is acceptable, and thus absolute effect sizes may more accurately represent the panorama as viewed by agents in this literature. Since simulations revealed that funnel plot asymmetry is unreliable when using

---

[6] The expectation is that there is a predictable relationship between the two such that distribution of points can be encompassed within an inverted funnel (which looks like an upright triangle). Specifically, one expects that observations that have a small standard error (i.e., are high up along the y-axis), should be close to the real effect size, since they are very precise. In contrast, studies that are imprecise (with a large standard error, and thus near the bottom of the triangle) are more noisy reflections of the underlying effect size and could be further away from it. Importantly, observations that are imprecise should spread as far away in a positive direction as they do in a negative direction – e.g., they should underestimate the effect size as often as they overestimate it. However, if a research field engages in selective reporting (i.e., if authors fail to submit, and/or editors to accept, certain results), then one may observe asymmetric funnel plots, usually sporting more effect sizes that overestimate the actual effect size than records that underestimate it.

absolute effect sizes, I instead used a correlation test to assess whether there was a relationship between the absolute effect size and the sample size. The estimate was negative as expected if there was selective reporting of larger absolute effect sizes, though not significant: $r(32) = -0.25$, $p = 0.154$. There was some evidence for asymmetry in the funnel plot based on absolute effect sizes, where the trim-and-fill method has added 6 extrapolated points near zero in an attempt to symmetrize, consistent with selective publication being informed by the size of the *absolute* effect size.

If absolute effect sizes were a better representation of underlying results, analyses for methodological moderators using absolute, rather than raw, effect sizes could lead to significant results in the moderator test. This expectation was not confirmed: As with systematically-coded effect sizes, the moderator test was not significant [$QM(6) = 7.275$, $p = 0.296$; see Appendix A.2, Table 6, for the precise estimates for each moderator]. Although in this analysis the beta for age reached, and that for contrast approached, significance, in each case the regression seemed to be pulled by one or a few data points, and in follow-up regressions declaring only one moderator (either age or contrast) the test for moderators failed to achieve significance (see https://osf.io/g4dc4/, pp. 9–13).

Our final analysis for bias involves the distribution of reported p-values that meet the threshold of significance. Twenty of the 34 unique infant groups were associated with significant results. The distribution of reported significant p-values (shown in Table 1) showed both a peak to the left, which is compatible with there being evidential value, and a peak to the right, consistent with p-hacking to meet the .05 threshold.

2.3. Discussion

A meta-analysis was carried out to integrate 34 independent effect sizes published in journals, representing the results of 570 infants, a sample size that is 71 times as large as the sample sizes found in the seminal study (Chambers et al., 2003). All of these experiments used similar methods to that original work. And yet a meta-analytic approach revealed that (a) the overall effect size was very close to zero; (b) proposed moderators did not significantly account for variance; and (c) there was weak evidence of bias in the literature. I discuss each in turn.

The overall effect size was not different from zero because in some cases results appeared to fit a novelty preference (with longer looking times to illegal than legal or familiar trials), and in others a familiarity preference. The uncertainty regarding the direction of preference that infants show is certainly not a new issue in infant research (e.g., Aslin, 2007; Hunter & Ames, 1988), and it has even been brought up in an extensive critique of the seminal infant phonotactic learning study (Weitzman, 2007). Although all infant researchers that use such preferential looking time methods are well aware of this uncertainty, our assumption is that these flips can be explained by scientifically-relevant factors. In other words, we believe that our methods are nonetheless experimental: If we took into account some additional factors, then preference flips could be predicted.

However, when I entered such factors in a meta-analytic regression, they did not account for a significant proportion of the variance. In other words, these explanations we proposed post hoc may well fit the individual data points included in the studies evoking them, but they have no predictive value with respect to the literature as a whole. This appears surprising given that the studies here fall within the categorization of 'laboratory experiments', where both exposure and test are tightly controlled by the experimenters. I will introduce possible interpretations of this state of affairs in the General discussion.

There were two more factors mentioned in the literature as affecting strength and direction of preference, but which were not be coded in the meta-analysis. As to the first, Gerken and colleagues (Gerken & Knight, 2015; Gerken & Quam, 2016) have recently shown that infants' preferences for legal versus illegal phonotactic

sequences can be affected by slight differences in the exposure stimuli. One group of infants in Gerken and Knight (2015)'s Experiment 1 heard 4 training exemplars that allowed infants to entertain a phonotactic rule at two levels of abstraction (individual sounds and classes of sounds), whereas the 4 training exemplars presented to the other half of the infants could only be interpreted at the level of the class. Both groups were tested with 8 legal and 8 illegal exemplars that were repeated for as long as infants looked. At this stage, infants in the former group showed a familiarity preference and those in the latter group a novelty preference. Thus, even slight differences in the materials, introducing alternative interpretations, could have remarkably long-lasting effects on infants' preferences. Might variance in phonotactic results meta-analyzed here be due to the chance presence of alternative interpretations in the training stimuli? Detailed inspection suggests that this may not be an issue, as several of the studies used a completely balanced set of stimuli, where legal and illegal, training and test items are perfectly balanced across infants (Chambers et al., 2003, 2011; Seidl et al., 2014). Thus, at least a sizable proportion of the studies here cannot be faulted with alternative interpretations when all training materials are considered.

However, additional work from the Gerken lab suggests that perhaps it is not the whole set of stimuli that matters, but a subset - and it is unclear how large of a subset. In Gerken and Quam (2016), two groups of infants were presented with the same set of 24 training items but in a different order; one group failed to show a significant preference, which the others did show. These authors suggested that infants may sometimes be drawing spurious conclusions from salient local co-occurrences, which at least in theory could explain some of the variance found in previous work. In practice, however, it is difficult to tell in which order stimuli were presented to each individual infant based on published methods. I suspect it might also be difficult to tell what precise generalizations can be drawn when many training words are used, since we do not know which items infants are latching on to in order to make their generalizations. As just mentioned, infants in Gerken and Knight (2015)'s study were able to draw generalizations from the 4 training items, and hold on to them while listening to 16 test items repeated for as long as infants looked. But infants could draw spurious generalizations from even smaller training sets. In fact, recent work suggests that infants exhibit a significant preference between two words depending on whether they have recently heard a single item that has low phonotactic frequency in the ambient language - showing that infants can generalize from a single type (Gerken, Dawson, Chatila, & Tenenbaum, 2015). These results suggest that it may be impossible to investigate certain phonotactic patterns, for instance one that bears on a class of sounds, as the effective training set actually used by the child might be the last item (or that with the lowest phonotactic probability in the input, or the highest pitch, …).

The possibility that the infant phonotactic learning literature suffers from bias was also empirically assessed using meta-analytic methods. There was no funnel plot asymmetry when systematically coded effect sizes were investigated. A re-analysis using absolute rather than raw effect sizes revealed a different picture, showing weak evidence of selective reporting. This result may indicate that, since absolute effect sizes are more commonly inspected, they may be more appropriate when looking for evidence of selective publication in infant research. Finally, the distribution of significant p-values supported both the presence of an underlying effect (resulting in very low p-values) and some p-hacking (resulting in p-values just under the $\alpha = .05$ threshold being over-represented).

3. Distributional learning

The second strand of literature relevant to infant learning of phonology in the lab is composed of studies where infants are exposed to a series of syllables which represent a distribution of acoustic correlates. This second, independent, research strand is analyzed using meta-

analytic methods similar to those described above for the infant phonotactic learning experiments.

3.1. Methods

All of the PRISMA materials and analysis scripts are available for download from https://osf.io/9zd2a/.

3.1.1. Paper identification

The paper pool was composed by combining a list of papers known to the author with systematic searches on pubmed, scholar.google.com, and a number of infant conferences. The detailed timeline and methods for study identification can be found on https://osf.io/7byw3/.

3.1.2. Paper selection

As in the phonotactic learning meta-analysis, I sought to include only near-replications of the seminal study by Maye et al. (2002). There are some papers that could not be included because they did not target the same learning mechanism. Specifically, some papers contain exposures to a syllable minimal pair that is associated with different wordforms or different facial expressions (Feldman, Myers, White, Griffiths, & Morgan, 2013; Teinonen, Aslin, Alku, & Csibra, 2008). Since learning presumably occurs via a different mechanism than clustering of the acoustic signal (e.g., through audio-visual association), these papers are not included in the present meta-analysis.

I did however allow for variation in exposure and test phases as long as they could be conceived as being conceptually similar. As for exposure, all distributional learning papers contained a bimodal and a control condition, of which there were three types. One type of control consisted of exposure to irrelevant stimuli (randomly varying tones); the remaining two employed the same continuum as the bimodally-exposed group but drawing from it as in a unimodal distribution (with a peak in the center) or a flat distribution (equally from all steps).

As for test phases, Maye et al. (2002) and several other papers presented infants with two types of trials, where the two sounds to be distinguished were alternating versus another where they were not alternating (i.e., only one token was presented). Maye et al. (2008), instead, habituated infants to one of the sounds to be distinguished, and then assessed dishabituation when the infant was presented with the other sound. Similarly, Wanrooij et al. (2014) used one sound as standard and another as oddball in an oddball detection paradigm using evoked response potentials (ERPs). All three test procedures can provide a measure of discriminability of the two sounds. In contrast, Cristià, McGuire, Seidl, and Francis (2011) presented infants with four types of trials, all of which contained two tokens, sampled from different regions in space. Since there is no straightforward way to map these four test trials into a repeated measure indexing sensitivity to the sound contrast, Cristià et al. (2011) was excluded.

In sum, the inclusion criteria were: (a) Participants are typically-developing children, between the ages of 0 and 36 months: (b) Participants experience a passive exposure in the lab, and are subsequently tested via any behavioral or non-behavioral method also in the lab: (c) The exposure is to a multimodal distribution of acoustic correlates or a control, the test assesses discriminability of the sound contrast.

After applying the inclusion criteria, there were 13 papers that could be considered: 6 journal articles (Liu & Kager, 2014; Maye et al., 2002, 2008; ter Schure, Junge, & Boersma, 2016; Wanrooij et al., 2014; Yoshida et al., 2010), 1 article in proceedings (Liu & Kager, 2011), 1 chapter in a collection (Capel, De Bree, De Klerk, Kerkhoff, & Wijnen, 2011), 1 unpublished manuscript (Cristia, 2011), and 4 sets of data that had not been included in public written reports but had been presented at conferences as posters or talks (Fennell, Hudon, & Spring, 2012; Pons, Mugitani, Amano, & Werker, 2006; Pons, Sabourin, Cady, & Werker, 2008).
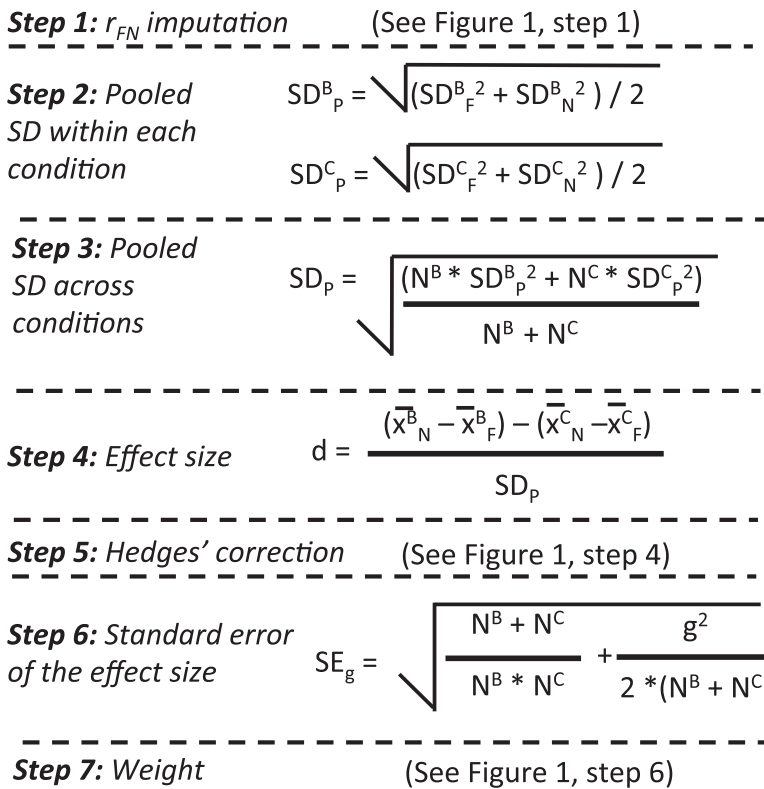
Step 1: r_{FN} *imputation* (See Figure 1, step 1)

Step 2: *Pooled SD within each condition*

$$SD^B_P = \sqrt{(SD^B_F{}^2 + SD^B_N{}^2)\,/\,2}$$

$$SD^C_P = \sqrt{(SD^C_F{}^2 + SD^C_N{}^2)\,/\,2}$$

Step 3: *Pooled SD across conditions*

$$SD_P = \sqrt{\frac{(N^B * SD^B_P{}^2 + N^C * SD^C_P{}^2)}{N^B + N^C}}$$

Step 4: *Effect size*

$$d = \frac{(\bar{x}^B_N - \bar{x}^B_F) - (\bar{x}^C_N - \bar{x}^C_F)}{SD_P}$$

Step 5: *Hedges' correction* (See Figure 1, step 4)

Step 6: *Standard error of the effect size*

$$SE_g = \sqrt{\frac{N^B + N^C}{N^B * N^C} + \frac{g^2}{2 *(N^B + N^C)}}$$

Step 7: *Weight* (See Figure 1, step 6)

Fig. 4. Procedure used to impute correlations and calculate standard deviations, effect sizes, standard errors, and weights. r_{FN} is the correlation in individual looking times during novel-like and familiar-like trials, with subscripts for each of the records (in light gray values that are not available); x_F is the average of looking times to less novel-like stimuli (non-alternating, habituation); x_N the average of looking times to novel-like stimuli (alternating, dishabituation); $SD_{F/N}$ their respective standard deviations and SD_P the pooled SD; N the number of infants included; d Cohen's d; g Hedges' g; SE_g the standard error of the effect size; w weight. The superscript B indicates that the estimate represents that for the bimodal group, and C that for the control group.

3.1.3. Data entry

The full list of fields entered can be found on https://osf.io/tynu9/. In addition to the same background variables as in the phonotactic meta-analysis (year, N of infants, mean age, etc.), for the present meta-analysis, I coded design (alternating/non-alternating versus habituation/dishabituation design, including the oddball ERPs study) and the sound contrast (consonantal versus non-consonantal, including both vowel and tone). Finally, I matched up each bimodal record with its control record.

3.1.4. Effect size calculations

Overall, there were 43 experimental records entered (26 drawn from journal publications), all of which had calculable effect sizes. Similar calculations used in the phonotactic learning meta-analysis were employed here, although with some additional complexity in that the present data compares performances across groups, and for a small subset of the data a different set of formulas were used. Most of the records corresponded to studies where a repeated measure was drawn within-participants (i.e., alternating and non-alternating, habituated category and new category, standard and oddball), with two groups of participants, one receiving a bimodal distribution and the other a control one. The procedure, represented in Fig. 4, is explained briefly in the next paragraphs.

Data were coded systematically with a single 'direction' of interpretation as follows. Looking times were coded using the same direction used in the phonotactic meta-analysis explained above, which is easy to apply to the habituation-dishabituation designs: Looking times during dishabituation trials map onto the more novel category, and looking times during habituation trials onto the more familiar type. As for the ERPs study, the mismatch response found had a positive polarity. Amplitude in response to the oddball was coded in the more novel column and amplitude in response to the standard in the less novel column. Finally, for alternating-non-alternating designs, it seemed reasonable to map the alternating looking times to the more novel, and the non-alternating to the less novel.

The first step, correlation imputation, is the same as used in the phonotactic meta-analysis (see Fig. 1, Step 1), except that I imputed correlations grouping studies based on design. For the 14 studies within the alternating-non-alternating design for which no correlation was available, one was imputed drawing with replacement from the 7 available ones for this design. For the 10 behavioral dishabituation studies with missing correlations, values were imputed drawing from the 4 available for this design.

In the second step, pooled SDs are calculated within each group just as in the phonotactic meta-analysis. In the third step, an overall pooled SD is then calculated across the bimodal and control-matched groups. To estimate the effect size (Step 4), one needs to do a 'difference of differences', which is divided by the standard deviation pooled across the two groups. This is followed by Hedges' correction exactly as in the phonotactic meta-analysis in Step 5. Next, the standard error of the effect size is calculated using the formula that is appropriate for between-group comparisons as Step 6. Finally, the weight is defined as the inverse of this standard error.

There was an exception to this general procedure, namely three papers for which difference scores and SD of the difference were available (Pons et al., 2006, 2008; ter Schure et al., 2016). The effect size was calculated as the ratio of the difference scores to the standard deviation, when the latter is corrected by a factor appropriate to the correlation in individuals' repeated measures.

3.1.5. Analyses

The present data set could have been analyzed in several different ways. For maximal clarity and conciseness, only one analysis is presented here; alternative ones are provided in the online supplementary materials. In the analysis presented here, effect sizes correspond to how much stronger the discrimination effect is in the bimodal infant groups compared to their matched controls. As in the phonotactic meta-analysis, therefore, one expects the intercept to be significantly higher than zero if the treatment (exposure to a bimodal distribution) actually improves infants' discrimination abilities in a consistent fashion. Also as in
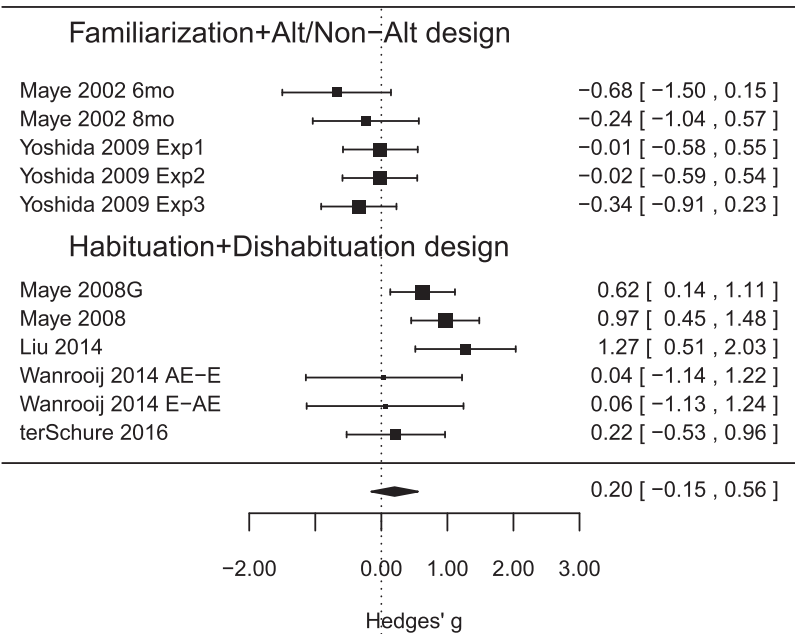
Familiarization+Alt/Non−Alt design

Maye 2002 6mo	−0.68 [−1.50 , 0.15]
Maye 2002 8mo	−0.24 [−1.04 , 0.57]
Yoshida 2009 Exp1	−0.01 [−0.58 , 0.55]
Yoshida 2009 Exp2	−0.02 [−0.59 , 0.54]
Yoshida 2009 Exp3	−0.34 [−0.91 , 0.23]

Habituation+Dishabituation design

Maye 2008G	0.62 [0.14 , 1.11]
Maye 2008	0.97 [0.45 , 1.48]
Liu 2014	1.27 [0.51 , 2.03]
Wanrooij 2014 AE−E	0.04 [−1.14 , 1.22]
Wanrooij 2014 E−AE	0.06 [−1.13 , 1.24]
terSchure 2016	0.22 [−0.53 , 0.96]

0.20 [−0.15 , 0.56]

−2.00   0.00   1.00   2.00   3.00

Hedges' g

Fig. 5. Forest plot of the sound category learning meta-analysis. Each row represents an effect size representing infants looking to novel-like trials minus familiar-like trials. The center of the square for that line indicates the observed effect size and its size represents the weight, whereas the whiskers delimit the confidence intervals. The gray diamond represents the overall effect.

Table 2

Results from a regression including age, exposure phase design, and type of contrast as potential moderators in the sound category meta-analysis. The beta indicates how much the effect size g changes with a given moderator; SE the standard error of that beta; z is the ratio between those two; and p the p-value for that ratio, rounded to three digits.

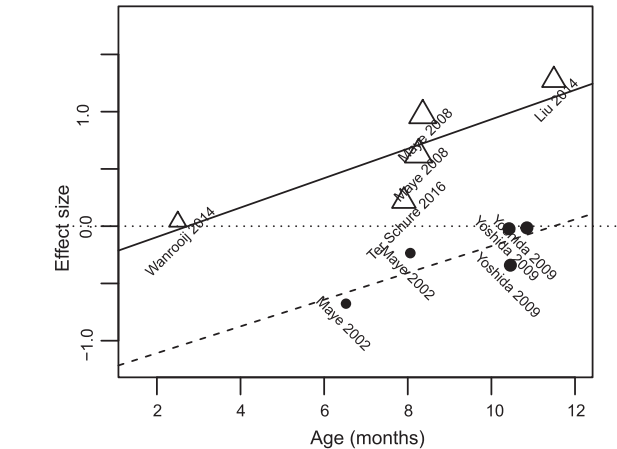| Moderator	beta	SE of the beta	z	p
Age	0.004	0.002	2.373	.018
Exposure phase	1.180	0.245	4.812	< .001
Contrast	0.159	0.294	0.541	.588

Fig. 6. Effect size as a function of age. Studies collected using the familiarization design are noted with circles, those using the habituation design with triangles. The size of the point represents the weight of the observation.

the phonotactic learning meta-analysis, I focus here on results including only on journal-published work (the meta-analysis including all work can be found on https://osf.io/w4q6h/).

To answer the second question, a meta-analytic regression is fit declaring conceptual and methodological dimensions as moderators. Three factors were thus investigated, the first one being the design (alternating-non-alternating versus habituation-dishabituation, with the latter including the ERPs oddball study). Additionally, two conceptual factors have been proposed to explain deviant results in specific studies, namely a linguistic factor (vowels are less amenable to

distributional learning than consonants; e.g., Pons et al., 2006); and age (older infants are less sensitive to distributional properties than younger ones; Yoshida et al., 2010).

Finally, the presence of bias was assessed exactly as in the meta-analysis above, by the inspection of funnel plot asymmetry and the distribution of reported significant p-values.

In the phonotactic meta-analysis, some analyses were repeated using absolute effect size instead of systematically-coded effects. There is no compelling rationale for doing the same here, since it appears unlikely that agents in the literature would ignore a change in direction of preference and instead compare size of preference across the two groups. Therefore, only systematically-coded effects are considered in this meta-analysis.

3.2. Results

Inspection of the distribution of the 26 journal-published effect sizes revealed that one data point in Wanrooij et al. (2014) was more than 3 standard deviations away from the mean over all studies, and was thus excluded. Additionally, data from the quiet-sleep and non-quiet sleep analyses in Wanrooij et al. (2014), constituting repeated measures, were averaged within each infant group. In all, there were 11 paired observations based on independent infant groups, whose results had been reported in journal articles (see Appendix B, Table 7).

A meta-analytic regression with no moderators revealed that the intercept was not significant, effect size g = 0.201, SE of the effect size g intercept = 0.181, z = 1.111, p = 0.266 (see forest plot in Fig. 5). The test for heterogeneity in effect sizes was highly significant, Q(10) = 29.046, p = 0.001, inviting further tests with moderators.

Following the planned analyses, an additional regression was fitted declaring age (centered), type of exposure phase design (familiarization versus habituation), and type of contrast (consonant versus non-consonant). Much of the inter-record variance was explained by exposure phase, with age also having a significant effect (see Table 2 for the precise estimates).

The effect of exposure phase design was explored by fitting additional regressions to the two subsets of data using different designs, including also age given its significant effect in the main moderator regression. For the 5 alternating-non-alternating paired records, the intercept was not significantly different from zero: effect size g = −0.486, SE of the effect size g intercept = 0.281, z = −1.729,
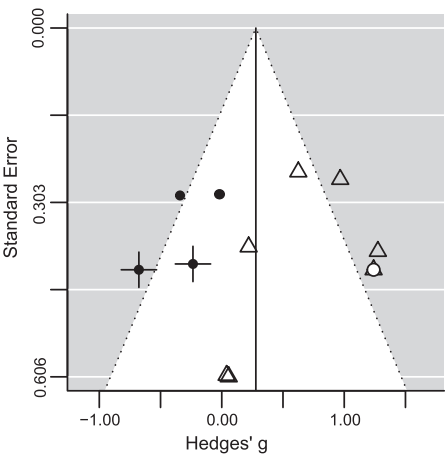
Fig. 7. Funnel plot of the sound category learning meta-analysis showing the standard error of the effect size as a function of the raw effect size for each pair of infant samples (data collected using the familiarization design shown with black circles, the habituation design with triangles). The white circle has been added by the trim-and-fill method in an attempt to symmetrize the distribution, and it happens to fall precisely over an extant habituation study. The two samples with an overlaid cross are those reported in the seminal study by Maye et al. (2002).

Table 3
Distribution of reported p-values below alpha = .05 in the literature on infant laboratory learning of sound categories.

(0, .01]	(.01, .02]	(.02, .03]	(.03, .04]	(.04, .05]
1	3	0	0	1

p = 0.084, and the estimator for age was not significant: β = 0.004, SE β = 0.003, z = 1.169, p = 0.243. The test for heterogeneity in effect sizes was not significant either, Q(3) = 0.957, p = 0.812.

For the 6 habituation-dishabituation paired records (including the ERPs study), the intercept is significant and positive: effect size g = 0.596, SE of the effect size g intercept = 0.149, z = 4.012, p < .001 and the estimator for age was also significant: β = 0.005, SE β = 0.002, z = 2.175, p = 0.030. The test for heterogeneity in effect sizes was not significant, Q(4) = 2.511, p = 0.643. Fig. 6 illustrates the effect of age for both groups of datapoints (those emerging from a familiarization exposure, and those emerging from a habituation exposure). Although the slope is markedly similar in two groups of data points, the interpretation is the opposite since, as a group, the preference approaches zero as infants age (starting from a negative origin) among studies using a fixed familiarization, whereas it becomes increasingly large (starting from a near-zero origin) among studies using a habituation design.[7]

Finally, the presence of bias was investigated. The funnel plot is shown in Fig. 7. A regression test for funnel plot asymmetry did not approach significance in the overall meta-analysis z = −0.718, p = 0.472; nor in the design-based subsets (among studies using a habituation exposure design z = −0.061, p = 0.951; among those using a familiarization exposure design z = 0.501, p = 0.617).

Finally, the distribution of reported p-values was more consistent with evidential value than p-hacking (see Table 3), although only 5 of the studies reported a significant difference between bimodal and control condition.

---

[7] Incidentally, this is the one analysis where the meta-analysis on journal-published versus all results diverge: When all results are included, age is not a significant predictor among familiarization studies. It remains significant among habituation studies, however.

3.3. Discussion

As with the meta-analysis on phonotactic learning, three key questions were addressed, namely the main effects, the importance of certain moderators, and the presence of bias, this time focusing on infant distributional learning of sound contrasts. When all journal-published studies that were conceptual replications and extensions of Maye et al. (2002) were included, the overall effect was not significant. However, the forest plot and subsequent analysis revealed this was due to clear heterogeneity stemming from a number of conceptual and methodological factors. Indeed, about half of the follow-up studies employed the original 2002 design, where infants are familiarized for a fixed amount of time and then presented with two types of trials at test, one in which the two categories to be learned alternate, and the other in which they do not. When only such studies were included, the overall effect size was not different from zero. The second group of studies followed a later paper, which used a habituation-dishabituation type of design to assess infants' discrimination of sounds along the acoustic dimension that had been manipulated during a prior exposure (Maye et al., 2008). Results from this second group of studies showed a significant advantage of bimodally-exposed infants compared to matched controls.

A moderator analysis confirmed that these two groups of studies diverged, and further confirmed that one of two previously proposed moderators (age, but not contrast type) explained a significant proportion of variance, although not necessarily in the proposed direction. Specifically, effect sizes increased with age within the dishabituation design, contrary to the proposal that older infants, who are beyond the 'perceptual reorganization' stage, should be increasingly resistant to such exposures.

Before moving on to the next section, I would like to draw some empirical consequences of the pattern of results found in the habituation-dishabituation subset, in order to inform future research. The overall effect size found in this subset was g = 0.596, which is a 'moderate' effect size by Cohen (1988)'s standards. This may give readers the impression that replicating such a result should be straightforward. Nonetheless, power analyses suggest that substantial sample sizes will be needed, as follows. Let us assume that future conceptual replications also aim to show that infants in the bimodal group do better than matched controls, and follow usual practices (e.g., set alpha to .05 and power to 80% following e.g. Cohen, 1992, using a one-sided test based on the expectation that bimodally-exposed infants cannot do *worse* than the control group) to do a prospective power analysis (Champely, 2009). This calculation suggests that we need about 35 infants in *each* group.

In consequence, future work should err on the side of cautiousness by keeping their sample sizes relatively large. There is a second argument supporting this same recommendation: Perhaps experiments that do not find sound category learning effects are less likely to be published in journals than studies that do, and therefore the overall category learning effect is smaller than what the present data reveals. While this is possible, there is not a great deal of evidence in this direction. To begin with, as noted previously, there is no trace of publication biases or p-hacking in this strand of work. Moreover, a meta-analysis including additional work not published in journals but appearing in chapters, theses, proceedings, and non-peer-reviewed venues yielded the same meta-analytic results (with one minor exception; see footnote 7 and Supplementary Materials https://osf.io/w4q6h/ pp. 6–7, for further details). Finally, extensive effort has been made to invite submissions through postings in mailing lists and direct contact of researchers known to be working on this topic, including through a form that can be submitted anonymously and requiring very little time from the person submitting it (live on 2016-12-19 at goo.gl/4zwtsc). Between July 2016 and June 2017, only two "file-drawer" replies were received, both using the fixed familiarization design followed by an alternating/non-alternating test. One reply described two experiments with a null result (no better performance in bimodal than control

infants) and the other described a single experiment with no significant preference among infants exposed to a bimodal condition. These non-significant effects fit in well with the meta-analytic results presented above, which suggest this design does not lead to reliable results.

4. General discussion

Several aspects of the results were common across the two meta-analyses. Regarding main effects, it was found that the body of work closely based on methods used in Chambers et al. (2003) and Maye et al. (2002) did not, as a group, confirm the original conclusions. The interpretation and implications of this fact are discussed in further detail below. Before moving on, it is important to remember that one pocket of stability was nonetheless found. A subset of studies, all of which used a design where only one direction of infant preference is interpretable (habituation-dishabituation), did provide statistical evidence for sound category learning. Additionally, it was found that previously posited moderators mostly did not explain significant proportions of variance, with the one exception behaving opposite to predicted (sound category learning effects *increased* with infant age within the dishabituation design). Finally, there was little trace of selective publication or inappropriate reporting practices, suggesting that, at least for this strand of research on infant cognition, researchers, reviewers, and editors do not engage in practices leading to a markedly biased literature. In consonance with this conclusion, analyses of all public literature (i.e., also including non-journal-published research) lead to essentially the same results as when analyzing only published research (see Supplementary materials for further detail). For this reason, the rest of the general discussion operates on the assumption that this is a faithful picture of actual results. In the remainder of this general discussion, I first lay out theoretical implications of these results, and then discuss a number of considerations related to the scientific study of infant cognition at large.

4.1. Theoretical implications for phonological acquisition

We found no systematic evidence that infants' *phonotactic* learning can be experimentally induced in the laboratory. This result may simply indicate that the field has not yet found a way to reproduce learning of sound patterns in a controlled setting, and with increased attention to stimuli design (for instance, by closely controlling how stimuli are built and presented to the infant, Gerken & Knight, 2015; Gerken & Quam, 2016), it may be possible to study this level of phonological acquisition *in vitro*. From this perspective, future research on phonotactic learning in infancy should study under what experimental conditions robust and replicable effects can be obtained, as having an empirical grasp on methods and results is a pre-requisite for any further work interpreting learning or failure to learn.

Alternatively, the overall null result could be viewed as a true negative: Perhaps the underlying phenomenon does not actually exist (for instance, if infants cannot compute phonological generalizations from experienced word forms, but instead must generalize from their long-term proto-lexicon), or perhaps phonotactic learning requires much longer exposures than those that can be provided in the laboratory. Some previous evidence is consistent with this interpretation. For instance, sensitivity to native phonotactics is not evident before 9 months of age (Jusczyk, Luce, & Charles-Luce, 1994), relatively late when compared to some other levels that may be resolved holistically (e.g., own name recognition, as in Mandel, Jusczyk, & Pisoni, 1995, word-object mapping for highly frequent and salient referents, as in Tincoff & Jusczyk, 1999). Interestingly, some research suggests that native phonotactics is more sensitive to length of exposure than maturation (Gonzalez-Gomez & Nazzi, 2012). One apparent counter-argument is that phonotactic learning in adults can sometimes be induced very rapidly (e.g., Onishi, Chambers, & Fisher, 2002); since the (arguably less plastic) adult system can nonetheless rapidly uptake

phonotactic information, it seems reasonable to suppose that infants should find it even easier to reshape their phonotactic expectations. This counter-argument, however, does not take into account the fact that adults presumably have much more robust sound category representations. It is conceivable that *rapid* phonotactic learning may need to await the development of robust sound categories. The suggestive fact that sensitivity to native language phonotactics roughly coincides with the timing of acquisition of certain consonantal contrasts has already been discussed elsewhere (Daland, 2009), another fact that fits in with the hypothesis that robust phonemic representations may be a pre-requisite to rapidly uptaking phonotactic information.

Turning now to sound category learning, the meta-analysis of distributional experiments revealed robust evidence for learning effects when considering a subset of studies that constituted a conceptual (albeit not a methodological) replication of the original Maye et al. (2002). The five journal-published studies in this body of data came from three different labs, established in two different countries, and thus represent the performance of infants routinely exposed to two different native languages (American English and Dutch). Thus, although this body of evidence is small, it is nonetheless promising in terms of the potential replicability of this line of research. It is particularly noteworthy that these experiments sampled from the whole range of phonological contrasts (consonants, vowels, and tones), suggesting that, under appropriate conditions, the mechanisms that shape sound discrimination based on brief exposures could be available to learn all of these category types. Moreover, although the studies spanned a considerable range of ages (two to 12 months), there was a trend for stronger results at older ages. From a theoretical viewpoint, then, we can conclude that the basic mechanisms whereby exposure to different distributions of acoustic cues shapes the discriminability of sound contrasts is available in infancy, although it may be most effective towards the end of the first year. Interestingly, this is also the time at which shifts in perception consonant with native language exposure have been observed, which could indicate that this process is sensitive to maturation (see a recent discussion in Werker & Hensch, 2015). Future work should apply a similar meta-analytic approach to the alternative sound learning strategies that have been proposed (as discussed in the Introduction; e.g., Feldman et al., 2013; ter Schure et al., 2016), and assess how infants select which strategy and/or whether they integrate several (e.g., both distributional and multimodal information).

4.2. Broader implications for infant cognition research

Despite the fact that experimental developmental psychology methods are relatively stable within a strand of research, the meta-analyses presented here, and those discussed elsewhere (Bergmann & Cristia, 2016; Cristia et al., 2014; Tsuji & Cristia, 2014) reveal above all the variability in results found across studies. This becomes obvious when adopting a systematic coding, where familiarity and novelty preferences are *not* equivalent, which seems epistemologically more appropriate. An experiment is defined as a set of conditions that are controlled in order to induce a measurable state. One result and the exact opposite, ensuing from essentially the same methods, cannot and should not be viewed as positive replications, even if both results are significant. Variance, particularly when reaching extreme cases entailing flips in preference, requires an explanation that allows us to predict the direction of results in a future study that follows essentially the same procedure. I meta-analytically explored the generalizability of several explanations that I and others had postulated ad hoc to account for apparently different results (flips of preference, lack of difference across groups), including age and linguistic differences (e.g., complexity of the phonotactic pattern, malleability of an acoustic dimension). None of the proposed mediators behaved quite as predicted.

There are three potential explanations for this state of affairs whereby proposed explanations do not hold up when all results are

taken into account. First, authors may postulate hypotheses to account for local results ad hoc but these hypotheses do not constitute general explanations in the sense defined above. Second, authors have found true effects and they have also put forward explanations with general validity to explain modulation of those effects, but both main effects and putative moderators have such small impact that we lack the meta-analytic power to pick up on them. Third, main effects and/or moderators are sizable, but they are currently invisible because there are other, unknown factors that have not been appropriately controlled. Although the latter two explanations are more charitable to the authors involved, they are nonetheless problematic from the point of view of the scientific study of infant cognition. They may mean that the measurement of infant cognition is so inaccurate that even when results from tens of studies are combined, to reach sample sizes of hundreds of participants, we still cannot detect crucial structuring variables. Or they could be taken to indicate that we are ignorant of key factors affecting infant preference, since we do not know what about the specific situation in which the data were collected and/or analyzed has such a great impact that it can flip the direction of preference or introduce such sizable variance.

4.2.1. Ambiguity in the causal pathway between exposure and response measured

To a certain extent, the variance may relate to the designs employed. In an open preference study, what is the direction of preference that indicates better performance? Let us consider first the case of alternating-non-alternating designs used in the distributional learning studies above (as well as in pure discrimination studies, Tsuji & Cristia, 2014), where arguments for a preference in both directions can be made with no clear ranking as to which should be better: The alternating trials are intrinsically more varied and therefore more interesting; and the non-alternating trials are more different from the familiarization and therefore more interesting. To say the least, this state of affairs is not amenable to increasing our comprehension of whether a finding is replicated or not, since we cannot even be certain what the base prediction is in terms of discriminability. In other words, the fact that the overall effect size for the alternating-non-alternating studies on distributional learning is not significant need not indicate that the distributional learning mechanism is unavailable; it could merely show that the relationship between experimental conditions and outcomes is poorly understood.

One may believe that the issue is less problematic for phonotactic learning, where a direction 'ranking' can be proposed. Indeed, a preference for illegal over legal trials can be viewed as stellar performance (with novelty indicating that infants fully incorporated the pattern); a preference for legal over illegal trials a less good but still acceptable performance (with familiarity indicating that infants discriminate the more familiar from the novel pattern even if they are not yet "bored" with it); and no preference as the worse. However, as has been argued previously on many occasions, the lack of preference is ambiguous, since it could arise from no learning but also from a transitional state between familiarity and novelty (e.g., Aslin, 2007; McMurray & Aslin, 2005). Notice that this issue is not entirely orthogonal to the problem raised in the discussion of the phonotactic meta-analysis, regarding the fact that infants may not uptake all the information the experimenter desired them to. In a sense, both issues relate to the fact that the link between the exposure intended for the infants and the resulting response is very indirect.

This ambiguity is not specific to studies on phonological learning. In fact, laboratory experiments have a fruitful history in the domain of infant cognition, as they appear as a useful tool to unveil the workings of learning mechanisms by isolating effects of a controlled exposure.

For instance, Younger and Cohen (1983) habituated infants to animal-like images in which certain features (e.g., number of legs and having feathers) correlated, and later tested for dishabituation when seeing a novel animal in which the correlation was maintained versus violated - a procedure that is extremely similar to the phonotactic learning studies meta-analyzed above. Similarly, Marcus, Vijayan, Rao, and Vishton (1999) exposed infants to 'words' which had a certain structure (e.g., 'leledi' represented an AAB abstract structure) for a fixed amount of time, and later presented novel words following and breaking that regularity. As clear from these descriptions, the main problematic features identified above could also apply to this work. In fact, the recent study showing generalization from a single type included an experiment in which infants generalized the AAB structure from a single word (Gerken et al., 2015). As a consequence, it would be extremely worthwhile to carry out meta-analyses on these other strands of infant cognition research, in order to assess whether the main effects found are more robust than those found here, and whether moderators postulated to explain 'deviant' results hold up as general explanations.

4.2.2. Limitations of dishabituation-based paradigms

Using paradigms that allow clearer predictions, such as habituation, can provide stronger constraints on our interpretations of empirical results. Nonetheless, there are a number of limitations that merit discussion. First, habituation/dishabituation may not be appropriate to all types of linguistic processing. In such cases, however, researchers could attempt to find paradigms that restrict the space of reasonable possibilities of interpretation. For instance, for wordform recognition, electro-encephalographic measures may be preferable over behavioral paradigms that can lead to the familiarity-novelty gray area (see Bergmann & Cristia, 2016 for a recent discussion).

Second, even when habituation/dishabituation is plausible, this is no assurance that results of multiple conceptual replications will be easy to integrate. A salient example of this was found in a recent cross-lab collaboration (Cristia, Seidl, Singh, & Houston, 2016), pooling together results of multiple independent experiments on test-retest reliability (i.e., the correlation in individual performance scores found when the same infants do a test twice). One subset of these experiments had been run in a single lab, drawing infants from the same population at roughly the same ages, and testing them on either the vowel contrast /i-u/ or the consonant contrast /s-ʃ/. These two experiments, despite their close twinning, yielded very different results. Infants tested on the vowel version exhibited a strong discrimination response on both test days (Hedges' g of 0.791 and 1.127 respectively), and individual performance across days was significantly correlated in the predicted positive direction (r = .31, p = .020). In stark contrast, the sibilant study evoked a small (and non-significant) discrimination response on the first day the infants were tested (effect size g = 0.11), but, counter to all expectations, the response reversed direction on the second day, with significantly longer looking to the habituated category than the novel category at test (g = −0.221). Even more puzzling, individual performance across days was *anti*correlated (r = −.23, p = .030). There is no reasonable explanation for the direction of preference observed on the second day or the negative correlation across days, and although many potential post hoc interpretations may be put forward, the main take-home message from such diverse results emerging from the same lab applying the same paradigm with different stimuli is that a great deal needs to be learned about infant performance in such lab studies.

Thus, better understanding and management of our experimental methods is crucial if infant experimental research is to move away from current piecemeal accounts, in which we are minimally asked to be self-consistent within a paper, and maximally to propose some ad hoc

explanations for divergences across studies. To advance in this path, the use of pre-registration for confirmatory research (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) more transparent practices of data collection, analysis, and publication (Nosek & Bar-Anan, 2012; Nosek et al., 2015), and collaborative research assessing sheer variability in results across labs (Frank et al., 2017) may help us producers of this literature gain a firmer grip on infant experimental data.

4.2.3. Suggestions for future ground-breaking work

The cornerstones of the present work were two papers that contained, in addition to their respective theoretical proposals, results from two groups of infants each. The total sample sizes were 16 in one case and 24 in the other; in both cases results had been gathered with a single method, in a single laboratory, and potentially by a single experimenter. The seminal studies, however, do not discuss these factors as potential limitations – an oversight that probably affects most journal-published articles. Yet, it is entirely reasonable that the first experiments on a topic may not be replicable, as new territory is being explored. For the same reason, any such new empirical result should be viewed with a certain dose of skepticism - a position that is currently not widespread. This is obvious not only in terms of the number of citations garnered by the seminal papers (e.g., according to scholar.google.com Maye et al., 2002 had been cited 980 times, and Chambers et al., 2003 220, by 2016-12-20), but also in the level of investment in follow-ups. As producers of this literature know all too well, infants are an expensive commodity, and the community has 'invested' 406 infants in distributional learning, journal-published studies, and 570 in phonotactic ones. When also including studies that are public but not published in journals, the number goes up to a total of 1522, and this is also not counting infants who are excluded and studies that have not been made public. How may we better inform potential consumers and producers of this literature, so that they can optimize their (and the community's) investment?

One obvious first response may be to ask authors of such papers to provide a replication, but I do not think this addresses the issue. As mentioned previously, both Maye et al. (2002) and Chambers et al. (2003) relied on data from two groups of infants, so they *did* include an internal replication. I propose that a format that is a hybrid between a regular paper and a registered report[8] may be optimal, and it would work as follows. Authors of a ground-breaking study would submit their theoretical proposal with its initial experimental support and receive provisional acceptance; a second *conceptual* (i.e., with different stimuli and if possible carried out by an independent laboratory) and fully pre-registered replication would be carried out and added, regardless of its result. The resulting paper would contain the theoretical proposal with its original proof of principle, as well as a second empirical data point. If the replication confirms the original results, this would suggest that the proposed mechanism is robust to methodological variation. If the replication does not confirm it, then the research community would still have access to the original proof and could further explore it in subsequent work, all the while being informed by both reported attempts. This proposal does not stifle exploration, since researchers are free to

develop and 'tweak' experimental paradigms until they find something that looks promising. At the same time, by requiring a single independent replication, the proposal lowers the cost for the community of checking whether the paradigm leads to predictable results, since it only 'costs' a second group of infants. Ideas similar to this one have already been explored in genome-wide association studies (Kraft, Zeggini, & Ioannidis, 2009), and more recently within psychology (see the Society for the Improvement of Psychological Science's "Study Swap" platform on http://improvingpsych.org/). Imagine what may have happened if the experiments in Maye et al. (2002) and Maye et al. (2008) had been published together: Although both ratified the predictions of the distributional learning theory, the results from Maye et al. (2008)'s paradigm were not only more easily interpretable, but also larger in sheer effect size, and thus they could have seemed more attractive for those carrying out follow-up work. As it turns out, their choice would have been justified, since results using this methodological paradigm turned out to be stabler in subsequent work than results following Maye et al. (2002)'s paradigm.

4.3. Conclusion

I have attempted to shed light on the strength of the empirical evidence supporting two putatively key mechanisms underlying infant language acquisition. Meta-analytic results highlighted the divergence in results found and the difficulty of postulating general moderators of the main effects, inviting reflection on the use of inherently ambiguous methods and potentially the stability of such results. One pocket of stability was found, namely it appeared that infants' perception of sound categories is affected by the distributions of acoustic cues in ways that may be systematically measurable using one specific type of design. Finally, there was only a little evidence of selective reporting or other publication biases.

---

[8] A common process for registered reports is as follows: First, a manuscript is submitted containing (a) a strong motivation for a research question; (b) a justification of an optimal way to answer the question with detailed explanations for methods, procedure, and analyses; and (c) an interpretation of potential outcomes, showing how they are all theoretically informative. Second, reviewers accept or reject the proposal (typically based on a), or provide comments to improve the proposal (usually addressing b and c). When an agreement has been achieved, the manuscript is accepted for publication regardless of what results are. Only then do the manuscript's authors collect the data. Finally, results and conclusions are added to the manuscript, which may be re-reviewed for clarity (see Chambers, 2015 for further information of the consequences of adopting registered reports for a research field). As clear from this description, the registered report is mainly appropriate for confirmatory work, where a great deal is known about the methodological paradigm and results can be neatly predicted. It is therefore not appropriate for ground-breaking studies, where some 'tweaking' of the methods is to be expected.

Appendix A

A.1. Study table for the phonotactic meta-analysis

Table 4.

Table 4

Final table of effect sizes on infant phonotactic learning, with each article identified by the first author name and the last two digits of the year (sorted by year; see main text for full references). Each line corresponds to an independent effect size (i.e., separate group of infants); a given article may contain multiple experiments on separate infant groups and thus contribute multiple independent effect sizes. For reasons of space, the names of individual groups within each article are as concise as possible, but should be sufficient to recognize the source of the data in the original paper. N indicates the number of infants included, ES the effect size, ES SE the standard error of the effect size, and p the p-value as reported in the original study, with ';' marking such analyses collapsing across infant groups (e.g., across two age groups). Age codes the average age (in months) infants had when tested; R, the type of rule they were exposed to (first or second order); tokens, the number of tokens they were familiarized with; types, the number of unique pseudowords they were familiarized with; and test indicates the type of trials against which illegal trials were being compared during test (familiar or legal).

Sample ID	N	ES	ES SE	p	Age	R	Tokens	Types	Test
Chambers03 1	8	0.85	0.40	.05	16.49	f	H (150)	L (25)	f
Chambers03 2	8	0.37	0.35	.05	16.49	f	H (125)	L (25)	f
Seidl05 1	24	0.42	0.18	.016	9.22	s	L (48)	H (48)	f
Seidl05 2	22	0.54	0.20	.006	9.22	s	L (51)	H (51)	f
Cristia08 1 FN	12	−0.38	0.29	.19	6.89	f	L (57)	H (57)	l
Cristia08 1 SN	12	0.48	0.21	.01	6.95	f	L (57)	H (57)	l
Cristia08 2 F	12	−0.59	0.30	.05;	6.85	f	L (26)	L (26)	l
Cristia08 2 S	12	−0.29	0.28		6.85	f	L (26)	L (26)	l
Seidl09 1	18	1.06	0.29	.006	11.09	s	H (120)	L (24)	l
Seidl09 2	36	−0.13	0.16	.18;	11.03	s	H (120)	L (24)	
Seidl09 3	36	−0.44	0.17		4.36	s	H (120)	L (24)	
Chambers11 1	16	0.48	0.25	.017	10.29	f	H (80)	L (16)	f
Chambers11 2 16.5m-g1	8	−0.71	0.38	.02	16.29	s	H (125)	L (25)	f
Chambers11 2 16.5m-g2	8	−0.99	0.42	.03	16.39	s	H (125)	L (25)	f
Chambers11 2 10.5m-g1	16	−0.16	0.24	.375	10.19	s	H (80)	L (16)	f
Chambers11 2 10.5m-g2	32	−0.38	0.17	.07	10.29	s	H (80)	L (16)	f
Chambers11 3 16.5m	16	−0.73	0.27	.007	16.19	f	H (80)	L (16)	f
Chambers11 3 10.5m	16	−0.72	0.27	.05	10.49	f	H (80)	L (16)	f
Chambers11 4 16.5m	16	−0.49	0.25	.01	16.19	f	H (80)	L (16)	f
Chambers11 4 10.5m	16	−0.36	0.25	.04	10.69	f	H (80)	L (16)	f
Seidl14 4mo-mult-	18	0.70	0.27	.003;	4.48	s	H (120)	L (24)	f
Seidl14 11mo-mult-	18	0.52	0.29		10.91	s	H (120)	L (24)	f
Seidl14 4mo-sing-talker	18	0.13	0.19		4.48	s	H (120)	L (24)	f
Seidl14 11mo-sing-talker	18	0.01	0.23		10.91	s	H (120)	L (24)	f
Wang14 1-8mo o	12	−0.19	0.33	.697;	7.88	f	L (60)	H (60)	f
Wang14 1-8mo c	12	0.01	0.38		7.88	f	L (60)	H (60)	f
Wang14 1-12mo o	12	0.36	0.16	.019	11.94	f	L (60)	H (60)	f
Wang14 1-12mo c	12	−0.46	0.29	.119	11.94	f	L (60)	H (60)	f
Wang14 2-15mo c	12	0.61	0.26	.026	14.97	f	L (60)	H (60)	f
Gerken15 1 pc	18	−0.50	0.24	.01	11.10	s	L (4)	L (4)	f
Gerken15 1 nc	18	0.52	0.24	.05	11.10	s	L (4)	L (4)	f
Gerken15 2	18	−0.13	0.23		11.10	s	L (4)	L (4)	f
Gerken16 1	20	−0.22	0.22		11.10	s	L (24)	L (24)	f
Gerken16 2	20	0.51	0.23	.04	11.30	s	L (24)	L (24)	f

A.2. Estimates for meta-analyses including moderators of phonotactic learning

Tables 5 and 6.

Table 5
Results from a regression including age, rule type, number of tokens and types, and type of test trial as potential moderators in the phonotactic meta-analysis using systematically-coded effect sizes. The beta indicates how much the effect size g changes with a given moderator; SE the standard error of that beta; z is the ratio between those two; and p the p-value for that ratio, rounded to three digits.

Moderator	beta	SE of the beta	z	p
Age	0	0.001	−0.374	.708
Rule	0.245	0.218	1.124	.261
Tokens	0.002	0.003	0.77	.442
Types-Low	−0.473	0.241	−1.965	.049
Types-Medium	−0.242	0.269	−0.9	.368
Test contrast	0.016	0.311	0.053	.958

Table 6
Results from a regression including age, rule type, number of tokens and types, and type of test trial as potential moderators in the phonotactic meta-analysis using absolute effect sizes. The beta indicates how much the effect size g changes with a given moderator; SE the standard error of that beta; z is the ratio between those two; and p the p-value for that ratio, rounded to three digits.

Moderator	beta	SE of the beta	z	p
Age	0.001	0.001	2.249	.025
Rule	0.075	0.111	0.677	.498
Tokens	0.001	0.001	0.963	.335
Types-Low	−0.017	0.113	−0.149	.882
Types-Medium	−0.061	0.129	−0.472	.637
Test contrast	0.317	0.162	1.959	.050

Appendix B. Study table for the distributional learning meta-analysis

Table 7.

Table 7
Table of effect sizes on sound category learning (sorted first by design, then by publication year; see main text for full references). Each line represents an independent (pair of) groups of infants. N stands for the number of infants included, ES for effect size, SE for the standard error of the effect size; these three numbers are available for the bimodally-exposed and control infants. Additionally, there is an effect size and SE of the ES for the contrast between bimodal and control conditions. Finally, p indicates the p-value associated with the effect of exposure (typically an interaction exposure x trial type) as reported in the original study, with ';' marking such analyses collapsing across infant groups (e.g., across two age groups). E codes the design of the exposure phase (h for habituation, f for familiarization), A the average age in months, and C the type of contrast (c for consonants, v for vowels, t for tones).

Cite	Bimodal N	ES	SE	Control N	ES	SE	Difference ES	SE	p	E	A	C
Maye02 6mo	12	−0.519	0.235	12	0.174	0.212	−0.677	0.420	.063;	f	6.5	c
Maye02 8mo	12	−0.345	0.221	12	−0.098	0.210	−0.235	0.410	NA	f	8.1	c
Yoshida09 Exp1	24	−0.030	0.148	24	−0.042	0.148	−0.015	0.289	.85	f	10.8	c
Yoshida09 Exp2	24	−0.108	0.099	24	−0.088	0.139	−0.022	0.289	.87	f	10.4	c
Yoshida09 Exp3	24	−0.232	0.082	24	0.083	0.167	−0.341	0.291	.018	f	10.4	c
Liu14	16	0.987	0.243	16	−0.195	0.185	1.275	0.388	.013	h	11.5	t
Maye08	32	0.681	0.153	32	−0.220	0.130	0.967	0.263	.001	h	8.4	c
Maye08G	35	0.344	0.129	32	−0.272	0.132	0.625	0.250	.05	h	8.3	c
terSchure16	13	0.319	0.234	15	0.022	0.187	0.219	0.380	NA	h	7.9	v
Wanrooij14 AE-E	6	0.113	0.215	6	0.410	0.509	0.040	0.603	NA	h	2.5	v
Wanrooij14 E-AE	6	0.139	0.264	5	0.039	0.157	0.057	0.606	.016;	h	2.5	v

References

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*(1), 48–53.
Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science, 19*(6), 901–917.
Bion, R. A., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLoS ONE, 8*(2), e51594.
Capel, D. J. H., De Bree, E. H., De Klerk, M. A., Kerkhoff, A. O., & Wijnen, F. N. K. (2011). Distributional cues affect phonetic discrimination in Dutch infants. *Sound and sounds. Studies presented to MEH (Bert) Schouten on the occasion of his 65th birthday* (pp. 33–43). Utrecht: UiL-OTS.

Chambers, K. E. (2004). *Phonological development: Mechanisms and representations* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction, 110*(1), 10–11.
Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87,* B69–B77.
Chambers, K. E., Onishi, K. H., & Fisher, C. (2011). Representations for phonotactic learning in infancy. *Language Learning and Development, 7*(4), 287–308.
Champely, S. (2009). pwr: Basic functions for power analysis. R package version 1.1.1. The R Foundation.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Lawrence Erlbaum.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Cristia, A. (2006). Speech sound categories in language acquisition and learning (Master's thesis). Purdue University. Retrieved from < https://sites.google.com/site/acrsta/nonresearch/CristiaThesis.pdf?attredirects=0 > .

Cristia, A. (2011). Infants' learning of tones and VOT contrasts in the lab. Unpublished manuscript. Retrieved from < https://osf.io/a486s/ > .

Cristia, A. (2015). Infants' ability to learn novel sound patterns changes with age. Unpublished manuscript. Retrieved from < https://osf.io/htsnp/ > .

Cristia, A., Seidl, A., & Gerken, L. (2011). Young infants learn sound patterns involving unnatural sound classes. University of Pennsylvania Working Papers in Linguistics, 17, Article 9.

Cristià, A., McGuire, G. L., Seidl, A., & Francis, A. (2011). Effects of the distribution of cues on infants' perception of speech sounds. *Journal of Phonetics, 39*, 388–402.

Cristia, A., & Peperkamp, S. (2012). Generalizing without encoding specifics: Infants infer phonotactic patterns on sound classes. *The 36th Annual Boston University Conference on Language Development (BUCLD 36)* (pp. 126–138). Cascadilla Press.

Cristia, A., & Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development, 4*, 203–227.

Cristia, A., Seidl, A., & Francis, A. (2011). Phonological features in infancy. In G. N. Clements, & R. Ridouane (Eds.). *Where do phonological contrasts come from? Cognitive, physical and developmental bases of phonological features* (pp. 303–326). Amsterdam & Philadelphia: John Benjamins.

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.

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

Dahl, D. B. (2009). xtable: Export tables to latex or html. R package version 1.7-0, retrieved from < http://CRAN.R-project.org/package=xtable > .

Daland, R. (2009). *Word segmentation, word recognition, and word learning: A computational model of first language acquisition* (Unpublished doctoral dissertation). Northwestern University.

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.

Fennell, C. T., Hudon, T., & Spring, M. (2012). Distributional learning of phonemes in monolingual and bilingual infants. *Presented at the 18th Biennial International Conference on Infant Studies*, Minneapolis, USA.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*.

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science, 18*(1), 80–89.

Gerken, L., & Knight, S. (2015). Infants generalize from just (the right) four words. *Cognition, 143*, 187–192.

Gerken, L., & Quam, C. (2016). Infant learning is influenced by local spurious generalizations. *Developmental Science, 20*(3), e12410.

Gonzalez-Gomez, N., & Nazzi, T. (2012). Phonotactic acquisition in healthy preterm infants. *Developmental Science, 15*(6), 885–894.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), e1002106.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128.

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

Ioannidis, J. P. (2005). Why most published research findings are false. *Chance, 18*(4), 40–47.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630–645.

Kraft, P., Zeggini, E., Ioannidis, J. P., et al. (2009). Replication in genome-wide association studies. *Statistical Science, 24*(4), 561–573.

Lewis, M., Braginsky, M., Bergmann, C., Tsuji, S., Cristia, A., & Frank, M. (2015). Metalab: A tool for power analysis and experimental planning in developmental research. *Presented at the Boston Conference on Language Development*, Boston, MA, USA. Retrieved from < http://metalab.stanford.edu > .

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

Liu, L., & Kager, R. (2011). How do statistical learning and perceptual reorganization alter Dutch infant's perception to lexical tones? *ICPhS, 17*, 1270–1273.

Liu, L., & Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition, 133*(2), 385–394.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science, 6*(5), 314–317.

Marcus, G., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by 7-month-old infants. *Science, 34*, 77–80.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*(1), 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can effect phonetic discrimination. *Cognition, 82*, B101–B111.

McMurray, R., & Aslin, D. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition, 95*, B15–26.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269.

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning. *Language and Linguistics Compass, 6*(11), 686–718.

Nosek, B. A., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., ... Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348*(6242), 1422–1425.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23*(3), 217–243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631.

Obrig, H., Mock, J., Stephan, F., Richter, M., Vignotto, M., & Rossi, S. (2017). Impact of associative word learning on phonotactic processing in 6-month-old infants: A combined EEG and fNIRS study. *Developmental Cognitive Neuroscience, 25*, 185–197.

Onishi, K., Chambers, K., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition, 83*, B13–B23.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius: Sociological Research for a Dynamic. World, 2*, 1–10.

Pons, F., Sabourin, L., Cady, J. C., & Werker, J. F. (2008). Distributional learning in vowel distinctions by 8-month-old English infants. *Presented at the 28th Annual Conference of the Cognitive Science Society*, Vancouver, BC, Canada.

Pons, F., Mugitani, R., Amano, S., & Werker, J. F. (2006). Distributional learning in vowel length distinctions by 6-month-old English infants. *International conference on infant studies, Kyoto, Japan.*

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from < http://www.R-project.org > .

Seidl, A., & Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development, 1*, 289–316.

Seidl, A., Cristia, A., Onishi, K. H., & Bernard, A. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development, 5*, 191–202.

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker variation aids young infants' phonotactic learning. *Language Learning and Development, 10*(4), 297–307.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547.

Smith, G. D., & Egger, M. (1998). Meta-analysis: Unresolved issues and future developments. *BMJ, 316*(7126), 221–225.

Sterling, T. D., Rosenbaum, W., & Weinkam, J. (1995). Publication decisions revisited. *The American Statistician, 49*(1), 108–112.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition, 108*(3), 850–855.

ter Schure, S., Junge, C., & Boersma, P. (2016). Semantics guide infants' vowel learning: Computational and experimental evidence. *Infant Behavior and Development, 43*, 44–57.

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

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

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*(33), 13273–13278.

Versteegh, M., Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The zero resource speech challenge 2015. *Proceedings of Interspeech.*

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638.

Wang, Y., & Seidl, A. (2015). The learnability of phonotactic patterns in onset and coda positions. *Language Learning and Development, 11*(1), 1–17.

Wanrooij, K., Boersma, P., & van Zuijen, T. L. (2014). Fast phonetic learning occurs already in 2-to-3-month old infants: An ERP study. *Frontiers in Psychology, 5*, 77.

Weitzman, R. S. (2007). Some issues in infant speech perception: Do the means justify the ends. *The Analysis of Verbal Behavior, 23*(1), 17–27.

Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Psychology, 66*(1), 173.

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition, 130*(1), 96–115.

White, K. S., Peperkamp, S., Kirk, K., & Morgan, J. (2008). Rapid acquisition of phonological alternations by infants. *Cognition, 107*, 238–265.

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in r. *Implement Reprod Res, 1*, 20.

Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy, 15*, 420–433.

Younger, B. A., & Cohen, L. B. (1983). Infant perception of correlations among attributes. *Child Development, 54*, 858–867.