A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building

Draft for Infancy, Special Issue, Due October 1

Michael C. Frank (Stanford University)
Elika Bergelson (Duke University)
Christina Bergmann (Ecole Normale Superieure)
Alejandrina Cristia (Ecole Normale Superieure)
Caroline Floccia (University of Plymouth)
Judit Gervain (CNRS & Université Paris Descartes)
J. Kiley Hamlin (University of British Columbia)
Erin E. Hannon (University of Nevada, Las Vegas)
Melissa Kline (Harvard University)
Claartje Levelt (Leiden University)
Casey Lew-Williams (Princeton University)
Thierry Nazzi (CNRS and Université Paris Descartes)
Robin Panneton (Virginia Tech)
Hugh Rabagliati (University of Edinburgh)
Melanie Soderstrom (University of Manitoba)
Jessica Sullivan (Skidmore College)
Sandra Waxman (Northwestern University)
Daniel Yurovsky (University of Chicago)

Abstract: The ideal of scientific progress is that we accumulate measurements and integrate these into theory, but recent discussion of replicability issues has cast doubt on whether psychological research conforms to this model. Developmental research – especially with infant participants – also has discipline-specific replicability challenges, including small samples due to high recruitment costs and limited measurement methods. Inspired by collaborative replication efforts in cognitive and social psychology, we describe a proposal for assessing and promoting replicability in infancy research: large-scale, multi-lab replication efforts aiming for a more precise understanding of key developmental phenomena. The ManyBabies project, our instantiation of this proposal, will not only help us estimate how robust and replicable these phenomena are, but also gain new theoretical insights into how they vary across ages, linguistic communities, and measurement methods. This project has the potential for a variety of positive outcomes, including less-biased estimates of theoretically-important effects, estimates of variability that can be used for later study planning, and a series of best-practices blueprints for future infancy research.

1. The "replication crisis" and its impact in developmental psychology

What can we learn from a single study? The ideal of scientific progress is that we accumulate progressively more precise measurements and integrate these into theories whose coverage grows broader and whose predictions become more accurate (Kuhn, 1962; Popper, 1963). On this kind of model, a single study contributes to the broader enterprise by adding one set of measurements with known precision and stated limitations. Then further studies can build on this, reusing the method and expanding scope, adding precision, and generally adding to a cumulative picture of a particular area of interest. Unfortunately, recent developments have cast doubt on whether research in psychology – and perhaps in other fields of inquiry as well – conforms to this model of cumulativity (Ioannidis, 2005; 2012). For example, a large-scale attempt to reproduce 100 findings from high-profile psychological journals found that fewer than half of these were successful (Open Science Collaboration, 2015; cf. Gilbert et al., 2016; Anderson et al., 2016).

There are many hypothesized sources for failures to replicate. First, any measurement includes some amount of random noise leading to variations in the outcome even if the exact same experiment is repeated under identical conditions (as happens, for instance, in a computer simulation; Stanley & Spence, 2014). When statistical power is low (e.g., through small sample sizes) this noise is more likely to lead to spurious, non-replicable findings (Button et al., 2013). Second, psychological findings are affected by many different contextual factors, and a poor understanding of how these factors affect our ability to measure behavior – whether in the experimental task, the test population, and/or the broader setting of the experiment – can cause a failure to replicate due to lack of standardization between researchers (Brown et al., 2014).

Third, a variety of "questionable research practices" can lead to improper statistical inferences. These include undisclosed analytic flexibility ("*p*-hacking"; Simmons et al., 2011) as well as other practices that bias the literature, such as the failure to publish null results (Rosenthal, 1979). All of these – and other – factors add up to create a context in which published experimental findings may inspire somewhat limited confidence (Ioannidis, 2005; Smaldino & McElreath, in press).

The research practices that limit replicability in the broader field of psychological research are present, and maybe even exacerbated, in developmental research. Developmental experiments often have low statistical power due to small sample sizes, which in turn arise from the costs and challenges associated with recruiting and testing minors. Consistent measurement of infants and young children is difficult because these participants have short attention spans and exhibit a large range of variability both within and across age-groups. Further, measures must rely on a small set of largely undirected responses to stimuli (e.g. heart rate, head-turns); direct instruction and explicit feedback are not possible in infancy research. In addition, young participants may spontaneously refuse to attend or participate during an experimental session. Due to this potential "fussiness," there are higher rates of data exclusion in developmental research than in adult psychology research; the need to specify fussiness criteria itself may also create further undisclosed analytic flexibility.

A related set of issues is tied to a general lack of methodological standardization: while many labs use similar methods, the precise setups vary, and there are few independent estimates of reliability or validity across labs (for discussion see e.g., Benasich & Bejar, 1992; Cristia et al., in press). Furthermore, initiatives that have incentivized replicability in other areas of psychology – preregistration, data sharing, and registered replication – have yet to become

widespread in the developmental community (although efforts for sharing observational data are a notable exception; cf. Adolph et al., 2012; Frank et al., 2016; MacWhinney, 2000; VanDam et al., 2016). This confluence of limitations may lead to replicability issues in developmental research that are more significant than currently appreciated.

Inspired by collaborative replication efforts in cognitive and social psychology (Klein et al., 2014; OSC, 2015), here we describe a proposal for assessing and promoting replicability in infancy research: large-scale, multi-lab replication efforts. The ManyBabies project, our instantiation of this proposal, aims to gain a more precise understanding of key developmental phenomena, by collecting data in a coordinated fashion across labs. These data will not only help us estimate how robust and replicable key phenomena are, but will also provide important new insights into how they vary across ages and linguistic communities, and across measurement methods. We believe this project has the potential for a variety of positive outcomes, including less-biased estimates of theoretically-important effects, estimates of variability (e.g., between labs or populations) that can be used for planning further studies and estimating statistical power, and a series of best-practices blueprints for future infancy research. In the remainder of the paper, we describe our approach and then go on to address some of the challenges of collaborative developmental work.

2. Collaborative data collection in infancy research

The aims of the ManyBabies project are importantly different from the aims of previous replication projects such as the Reproducibility Project: Psychology (OSC, 2015), which focused on estimating the replicability of an entire scientific field. Instead, our aim is to understand why different developmental labs, studying the same developmental phenomena using the same or

highly similar methods, might find differences in their experimental results. To achieve this goal, we plan to conduct a series of pre-registered, collaborative, multi-site attempts to replicate theoretically-central developmental phenomena. Thus, our approach is much more closely aligned with the "Many Labs" projects, from which we take our name. The Many Labs effort focuses on understanding variability in replication success and identifying potential moderators (e.g., Klein et al., 2014). But the effort involved in reproducing even one infant result across a large group of labs is substantial. To make the most of this effort and create high-value experimental data sets, we must navigate the tension between standardization across labs (with the goal of eliminating variability) and documentation of variability (with the goal of analyzing it).

For example, there is wide variation in experimental paradigms implemented across infant labs, manifest in both the paradigms that are available in a given lab and in how these paradigms are implemented. For practical reasons, it is not possible to use a single identical paradigm across labs, so in the ManyBabies 1 study described below, we will include several standard paradigms for measuring infant preferences (habituation, headturn preference, and eye-tracking). Each lab using a particular paradigm will be provided with a collaboratively-developed protocol to minimize within-paradigm variability. Deviations from these standards within individual labs, where necessary, will be carefully documented.

As a second example of the tension between standardization and documentation, it is clearly impossible to standardize all aspects of the sample of infants that we recruit across sites. Instead, we will document participant-level demographics (e.g., native language, mono- vs. bilingual environment, socio-economic status). In general, our approach will be to choose a

relatively small set of potential lab- and participant-level moderators of experimental effects in each project and plan analyses that quantify variation on these variables.

In addition to those sources of variation that can be straightforwardly documented and analyzed, there will be other systematic variation across labs on dimensions that are more difficult to quantify, like physical lab space, participant pool, and experimenter interaction. One of the goals of the project is to measure the variability in effect size that emerges from such sources, which is typically difficult to separate from truly random variation. Minimally, we will be able to make precise estimates of the proportion of variance that is explained by (structured) lab-to-lab variation. With the hope of potentially exploring ultimate sources of structured between-lab variation, the group is discussing supplemental steps we can take to ensure high data collection standards, including the video recording and sharing of all experimental procedures (e.g., using sharing platforms like Databrary; Adolph et al., 2012) and the training of RAs and other experimenters with standard videos across sites.

Because participant exclusion criteria, preprocessing steps, and specific analytic statistics all present opportunities for analytic flexibility (and hence an inflation of false positives), we will fix these decisions ahead of time. We will use both simulated and real pilot data to establish a processing pipeline and set standards for data formatting, participant exclusion, and the myriad other decisions that must be taken in data analysis. Once analytic decisions are finalized, we will pre-register our experimental protocol and analyses, freezing these confirmatory analyses (providing a model "standard operating procedure" for future analyses). This pre-registration does not, however, preclude exploratory analyses, and we anticipate that these will be a significant source of new insights going forward. In this spirit, all of our methods, data, and analyses will be completely open by design. We will use new technical tools (e.g., the Open

Science Framework) to share the relevant materials with collaborators and other interested parties. We hope this openness provides other unanticipated returns on our invested effort as others use and reuse our stimuli, protocols, data, and analysis code.

Having established a set of goals and an approach, our group next converged on a target case study. After an open and lively discussion with interested labs, we elected via majority vote to examine infants' preference for speech directed to them (infant-directed speech, or IDS) in our first ManyBabies replication study (MB1), described below. We decided to begin with an uncontroversial and commonly-replicated finding so as to provide some expectations for variability across labs and to provide guidelines for planning further studies. Indeed, further down the line, we hope to consider replications of a range of developmental phenomena, including both fundamental phenomena whose replicability is not in question as well as more controversial findings. We also recognize that there is no single approach to collaborative replication that will apply in all cases. For example, when attempting to replicate controversial findings, tight standardization will typically be necessary. However, attempts to assess the generalizability of a well-established finding will instead benefit from documenting variability. In sum, across many different possible targets, we believe that the collaborative approach will yield new empirical and theoretical insights.

3. ManyBabies 1: The preference for infant directed speech

Infants' preference for speech containing the unique characteristics of so-called infant-directed speech (IDS) over adult-directed speech (ADS) has been demonstrated using a range of experimental paradigms and at a variety of ages (e.g., Cooper & Aslin, 1990; Cooper, Abraham, Berman, & Staska, 1997; Fernald & Kuhl, 1987; Hayashi, Tamekawa, & Kiritani, 2001;

Newman & Hussain, 2006; Pegg, Werker & McLeod, 1992; Werker & McLeod, 1989). Moreover, infants perform better in language tasks when IDS stimuli are used, such as detecting prosodic characteristics (Kemler Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989) or learning/recognizing words (e.g. Singh, Nestor, Parikh, & Yull, 2009; Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011; Thiessen, Hill, & Saffran 2005). A typical experimental operationalization of a preference for IDS is that infants will attend longer to a static visual target (e.g., a checkerboard) when looking leads to hearing IDS, as opposed to ADS (Cooper & Aslin, 1990).

The preference for IDS is observed robustly across studies, but it is also quite variable. Data from a recent meta-analysis examining 34 published studies, including 840 infants (Dunst, Gorman, & Hamby, 2012) reveals significant heterogeneity [$Q(49) > 222$] and some evidence of publication bias ($z = 2.5$, $p = 0.01$; see Figures 1 and 2; data available at http://metalab.stanford.edu). In addition, while several moderators of the size of infants' IDS preference have been described (e.g., age), considerable variance remains unexplained. Most notably, although the presence of IDS is a cross-linguistic phenomenon (see Soderstrom, 2009 for review), there is variation across languages, and North American English appears to provide an especially exaggerated form (Fernald et al. 1989; Floccia et al., 2014; Shute & Wheldall, 1995; Kitamura, Thanavishuth, Burnham, & Luksaneeyanawin, 2002; though c.f. Farran, Lee, Yoo, & Oller, 2016). Although studies have found preferences for IDS in other languages, including Japanese (Hayashi et al., 2001) and Chinese (Werker et al., 1994), the tendency for studies on IDS to come from North America may therefore provide an inflated view of its robustness.

We selected IDS preference for our first replication study because it satisfies a number of key desiderata. Most importantly, it allows us to measure inter-lab and inter-baby variability

because the effect itself is large and robust – at least within North America. With a large effect as a baseline, we can assess variation in that effect across methods (e.g., comparing eye-tracking versus human-coded procedures), across linguistic communities (e.g., comparing infants for whom the stimuli are native versus not), and across ages. Distinguishing these moderators would not be possible in the case of a phenomenon with a smaller effect size. In the worst case, if the original effect were truly null, no moderation relationships would be detectable at all.

Preference for IDS also allows MB1 to assess several questions that are important for developmental theory. First, some views of language acquisition attribute a key role to IDS preference in scaffolding language learning, due to its attention-driving properties (e.g., Kaplan, Goldstein, Huckeby, Owren, & Cooper, 1995) or specific linguistic characteristics (e.g., Kuhl et al., 1997). Although the preference for IDS is robust with young infants, fewer studies precisely examine how this preference changes across development (although c.f. Hayashi et al., 2001; Newman & Hussain, 2006). Second, our study also offers an opportunity to examine the classic theory of native-language phonological specialization – in which general preferences and perceptual abilities gradually become language-specific over the course of the first year (Kuhl, 2004; Werker & Tees, 1984) – in a new domain. Use of IDS varies across linguistic communities (see above), but there has been relatively little study of this variation. For example, as mentioned above, British English IDS has less prosodic modification than North American English IDS (Fernald et al., 1989; Shute & Wheldall, 1995). Does this imply that UK infants might be particularly interested (or uninterested) in the intonational characteristics of North American IDS? Should this interest decline as they recognize that the dialect of the IDS is distinct (Butler et al., 2010)? We can ask the same question for infants learning languages other than English**:** will the IDS preference decline more quickly with age for these infants compared with North

American English (NAE) learning infants? In sum, the MB1 study provides both methodological and conceptual opportunities.

To address all of these questions, after substantial consideration, we elected to use precisely the same speech stimuli across labs: samples of IDS from NAE-speaking mothers, the most well-studied IDS source. Our study therefore measures preference for NAE IDS, rather than IDS more generally. This choice was a necessary compromise to meet our other goals. For example, if each lab recorded its own stimuli, lab-related variability would have been confounded with stimulus variability. Under this design, the burden of time and expertise for participating labs would have been substantially greater. We recognize that this decision furthers the existing NAE bias in the literature (Henrich, Heine, & Norenzayan, 2010), but since our goal was to replicate an existing phenomenon, we were constrained by the same literature. Our hope is that this initial study will spur additional research using other languages.

At the time of writing, we have divided into committees who are working towards making informed key decisions on all subsequent aspects of this project, including stimuli selection, experimenter training, data collection, and data analyses. We expect to begin data collection in late 2016 and complete it roughly a year later, analyzing and writing up the main results shortly thereafter. We also hope many more manuscripts from this project emerge as participants and the community more generally explore the resulting rich dataset.

4. Challenges and benefits of collaborative data collection

Any project has costs – in time, money, and research effort – and for a project as large as ManyBabies, these will be considerable. Nevertheless, we believe that the benefits of the project outweigh these costs, both for individual researchers and for the field as a whole. We discuss

challenges and benefits for individuals (including early career researchers), labs, and the field as a whole.

4.1 Individual researchers

One obstacle to collaborative projects like ManyBabies is that the positive incentives for participation are not as obvious as those for independent research. For example, grant panels and tenure committees may be strongly focused on first- and last-author publications, and may not sufficiently recognize collaborative work even when specific contributions are carefully documented. Early career researchers (ECRs) in particular are especially vulnerable to the need to produce original scholarship on a relatively short timeline. But given the relatively modest investments of time and effort necessary to make a contribution to a large project, we believe these potential downsides are outweighed by a number of substantial positive benefits.

Improvements in individual scientific practices. Issues of replication and reproducibility are fundamentally not just problems for the community as a whole, but also problems for individual researchers who may both fail to perform replicable research and fail to replicate others' work. Collaborative projects allow individual researchers to gain experience with community-generated best-practices in experimental design, data analysis, and use of collaborative open-science tools. Such opportunities may be especially valuable for ECRs who do not have access to local training in these practices. As a group, the authors of this paper have found the discussions surrounding project planning to be helpful with their own evolving understanding of issues of reproducibility and study design.

Being a pioneer. Although there are still significant impediments, attitudes towards the value of collaborative and replication work are changing. In the coming years, contributions to collaborative work and projects that work to resolve the replicability crisis may be important

factors in hiring, promotion, and funding decisions. Researchers who can show a pattern of early adoption of these new attitudes and approaches will demonstrate a visible and potentially field-shaping commitment to replicability in psychological science.

Opportunities for secondary analysis. Large-scale collaborative projects yield a multitude of data that, in addition to the planned analyses, can be explored for different kinds of research questions, creating additional publication opportunities for the same effort.

Being part of a community. A final important component of the collaborative approach for its participants is the opportunity to collaborate with other researchers. Collaborative efforts provide significant opportunities for networking, mentorship, and the sharing and cross-fertilization of ideas, well beyond those afforded by the standard conference and publication paradigm. With the widespread use of videoconferencing, collaborative projects bring together researchers across timezones in relatively intimate, friendly, supportive, and significant interactions. For ECRs, collaborative projects provide a method for connecting with a broad community of interest and raising awareness about their own skills and abilities. In addition, connections made through collaborative projects may blossom into other professional interactions.

4.2 Groups, labs, and lab heads

Even if individuals may be interested in a collaborative project, the decision to commit the resources of a research group or lab may be more complex. For example, often labs have funding obligations that require a specific amount of administrative or participant recruitment resources to be devoted to ongoing projects. In the short term, the ManyBabies group has secured modest funding to support lab involvement where it would otherwise not be possible, but longer-term financial support may be important for sustaining the group's efforts. But again, as in the case of

individual researchers, there may be a variety of other subsidiary benefits that outweigh the costs of participation.

Standardization of research practices with other labs. In group discussions regarding the standardization of practices across labs for ManyBabies, many previously-unrecognized differences in lab practice have emerged (e.g., deciding what counts as 'piloting' or when it is acceptable to restart an experimental session). Understanding how different sources of variability impact the robustness and replicability of experimental effects may help labs improve their own practices. Furthermore, in the long term, new laboratories could use a ManyBabies protocol to calibrate their laboratory and compare their data against the group standard as a means of establishing reliability.

Implementation of emerging open science practices. The ManyBabies project makes use of a number of emerging practices to ensure reproducibility and to facilitate communication and dissemination, as discussed above. These practices – for example, creating shared project repositories, generating analysis or simulation pipelines, and writing pre-registration documents – provide the same kind of benefits to efficiency and reproducibility when used within a single lab. Contributors to the Many Babies study will be able to bring these tools back with them to their home lab.

4.3 The field as a whole

Finally, we believe that there are important reasons why collaborative projects benefit the field as a whole. Replication work leads to more robust science, greater confidence in our findings, and better knowledge-sharing about methodological concerns. We highlight two such important consequences here.

Funding is tied to community confidence. As any researcher knows, the public controls the purse strings – if a government is voted into office that is less friendly toward research (or even certain kinds of research), this decision will be very quickly felt within the research community by individual researchers who do not get the grant funding they rely on for their work. Setting aside altruistic desires to do high-quality research, our own self-interest in the field of developmental psychology – and infancy research in particular – should drive us to support endeavors like ManyBabies in order to demonstrate to the public our commitment to improving scientific practice and outcomes.

Creating "best practice" materials and guidelines for experimental procedures and data analysis. The first ManyBabies project will create a push-and-play implementation of a discrimination experiment with a directional prediction. The natural side effect of this study is that it will lead to a set of consensus decisions about experimental procedures using different paradigms for measuring preference, and a set of open-source analysis scripts for the kind of data such experiments generate. These materials will not only lower commitment costs for labs involved in the study, but will also create a well-realized template for any future work (by ManyBabies participating labs or otherwise) wanting to use these popular developmental methods.

5. Conclusions

The foundational purpose of developmental research is to create and disseminate knowledge about organic changes in learning and behavior over age. This goal is best achieved within a culture of careful, methodological research. The ManyBabies project is a new collaborative effort to promote best practices, evaluate and build on influential findings, and understand

different dimensions of variability in laboratory-based infancy research. While data collection is currently ongoing for our first project on infant-directed speech preference, other tangible benefits have already emerged from this collaboration.

First, the ManyBabies project served as one inspiration for a recent pre-conference at the International Congress on Infant Studies, titled *Building Best Practices in Infancy Research*. This pre-conference in turn triggered discussions with the Congress leadership, leading to the special issue you are now reading. Such "knock-on" effects are one important benefit of so many people's efforts being directed at these issues.

Second, participation in the ManyBabies project has already affected how we conduct our own research. For instance, there have been many fruitful discussions during video conferences among members of the ManyBabies 1 project subgroups, including the methods, stimulus, data analysis, and ethics groups. Both macro-level conceptual issues about conducting rigorous research and micro-level methodological issues about a variety of topics have been discussed, such as how to most effectively reduce parent interference during experiments. These discussions have already informed practices within our own labs.

Third and finally, the project has served to promote community-building in infancy research outside of the standard framework. This venue for interaction is likely to enhance the rigor and health of our field in the future by promoting reproducibility, improving methods, sharing ideas and data, encouraging reasonable interpretations of data, and building theories. Central to this effort is the idea that science is fundamentally incremental and collaborative. All graduate students, postdoctoral researchers, research staff, and faculty members are welcome to join the ManyBabies project and our community more generally.

References

Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry*, *23*, 244-247.

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., ... & Della Penna, N. (2016). Response to comment on "estimating the reproducibility of psychological science". *Science*, *351*, 1037.

Benasich, A. A., & Bejar, I. I. (1992). The Fagan test of infant intelligence: a critical review. *Journal of Applied Developmental Psychology, 13*, 153-171.

Brown, S. D., Furrow, D., Hill, D. F., Gable, J. C., Porter, L. P., & Jacobs, W. J. (2014). A Duty to Describe Better the Devil You Know Than the Devil You Don't. *Perspectives on Psychological Science*, *9*, 626-640.

Butler, J., Floccia, C., Goslin, J., & Panneton, R. (2011). Infants' discrimination of familiar and unfamiliar accents in speech. *Infancy*, *16*, 392-417.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*,*14*, 365-376.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (in press). Test–Retest Reliability in Infant Speech Perception Tasks. *Infancy.*

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*,*1*, 4-17.

Cooper, R. P., Abraham, J., Berman, S., & Staska, M. (1997). The development of infants' preference for motherese. *Infant Behavior and Development, 20*, 477-488.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61*, 1584-1595.

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*, 1-13.

Farran, L. K., Lee, C. C., Yoo, H., & Oller, D. K. (2016). Cross-Cultural Register Differences in Infant-Directed Speech: An Initial Study. *PloS one*, *11*, e0151518.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16*, 477-501.

Fernald, A., & Kuhl, P. K. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development, 10*, 279-293.

Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., ... & Vihman, M. (2016). British English infants segment words only with exaggerated infant-directed speech stimuli. *Cognition*, *148*, 1-9.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the Reproducibility of Psychological Science". *Science*, *351*, 1036.

Hayashi, A., Tamekawa, Y., & Kiritani, S. (2001). Developmental change in auditory preferences for speech stimuli in Japanese infants. *Journal of Speech, Language, and Hearing Research, 44*, 1189-1200.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83.

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

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645-654.

Kaplan, P. S., Goldstein, M. H., Huckeby, E. R., Owren, M. J., & Cooper, R. P. (1995). Dishabituation of visual attention by infant-versus adult-directed speech: Effects of frequency modulation and spectral composition. *Infant Behavior and Development, 18*, 209-223.

Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language, 16*, 55-68.

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2002). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development, 24*, 372-392.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social Psychology, 45*, 142-152.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., ... & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*, 684-686.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Lloyd-Fox, S., Papademetriou, M., Darboe, M. K., Everdell, N. L., Wegmuller, R., Prentice, A. M., ... & Elwell, C. E. (2014). Functional near infrared spectroscopy (fNIRS) to assess cognitive function in infants in rural Africa. *Scientific Reports*, *4*, 4740.

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant-and adult-directed speech. *Language Learning and Development, 7*, 185-201.

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

Pegg, J. E., Werker, J. F., & McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development, 15*, 325-345.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638-641.

Scott, K. M. and Schulz, L. E. (in press). Lookit: a new online platform for developmental research. *Open Mind.*

Shute, B., & Wheldall, K. (1995). The incidence of raised average pitch and increased pitch variability in British 'motherese' speech and the influence of maternal occupation and discourse form. *First Language, 15*, 35-55.

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*, 1359-1366.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of Infant-directed speech on early word recognition. *Infancy, 14*, 654-666.

Smaldino, P. E. & McElreath, P. (in press). The natural selection of bad science. *Royal Society Open Science*.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic?. *Perspectives on Psychological Science*, *9*, 305-318.

Stern, D.M., Spieker, S., Barnett, R.K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language, 10*, 1-15.

Thiessen, E.D., Hill, E.A., & Saffran, J.R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7*, 53-71.

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

Figures

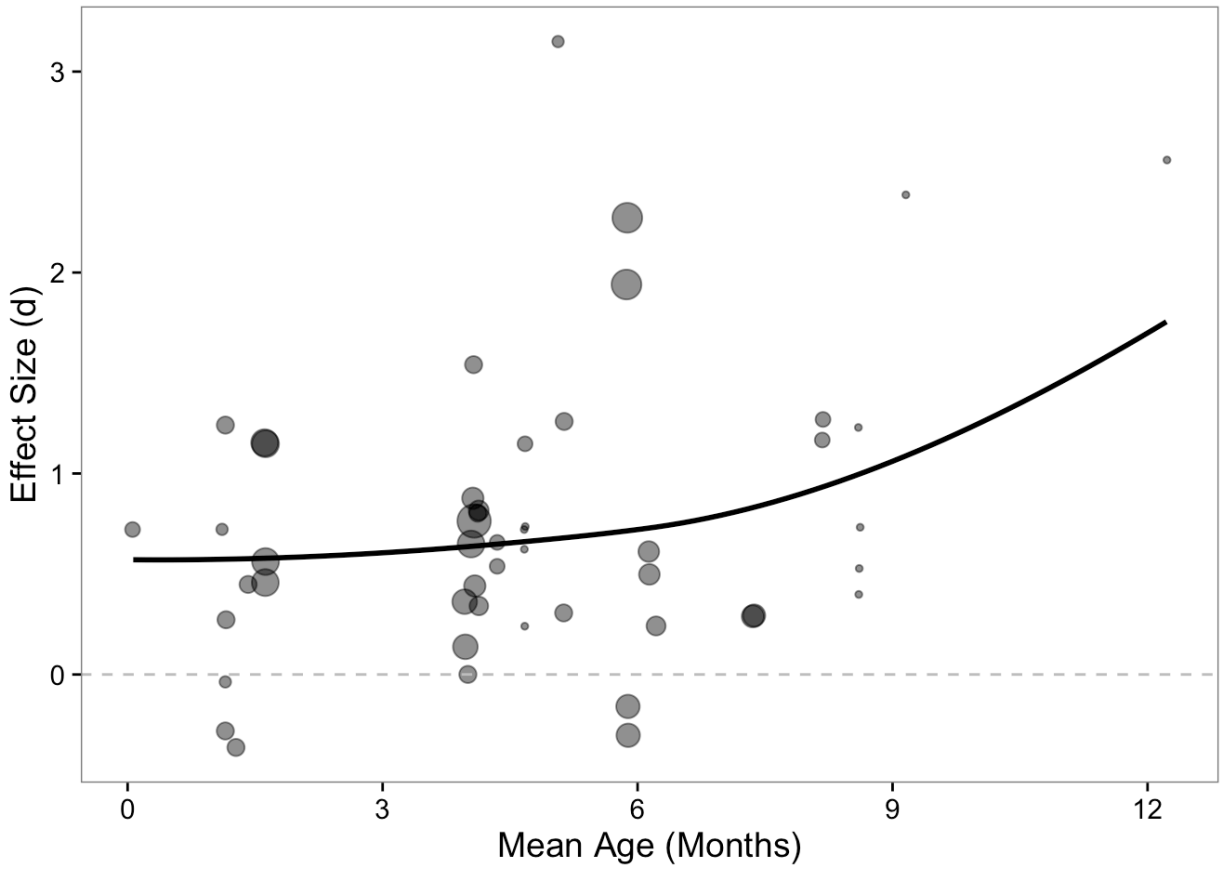

Figure 1. Meta-analysis of IDS preference, modified from http://metalab.stanford.edu. Points show individual studies, with point size showing N. Line shows an inverse-variance weighted local regression.

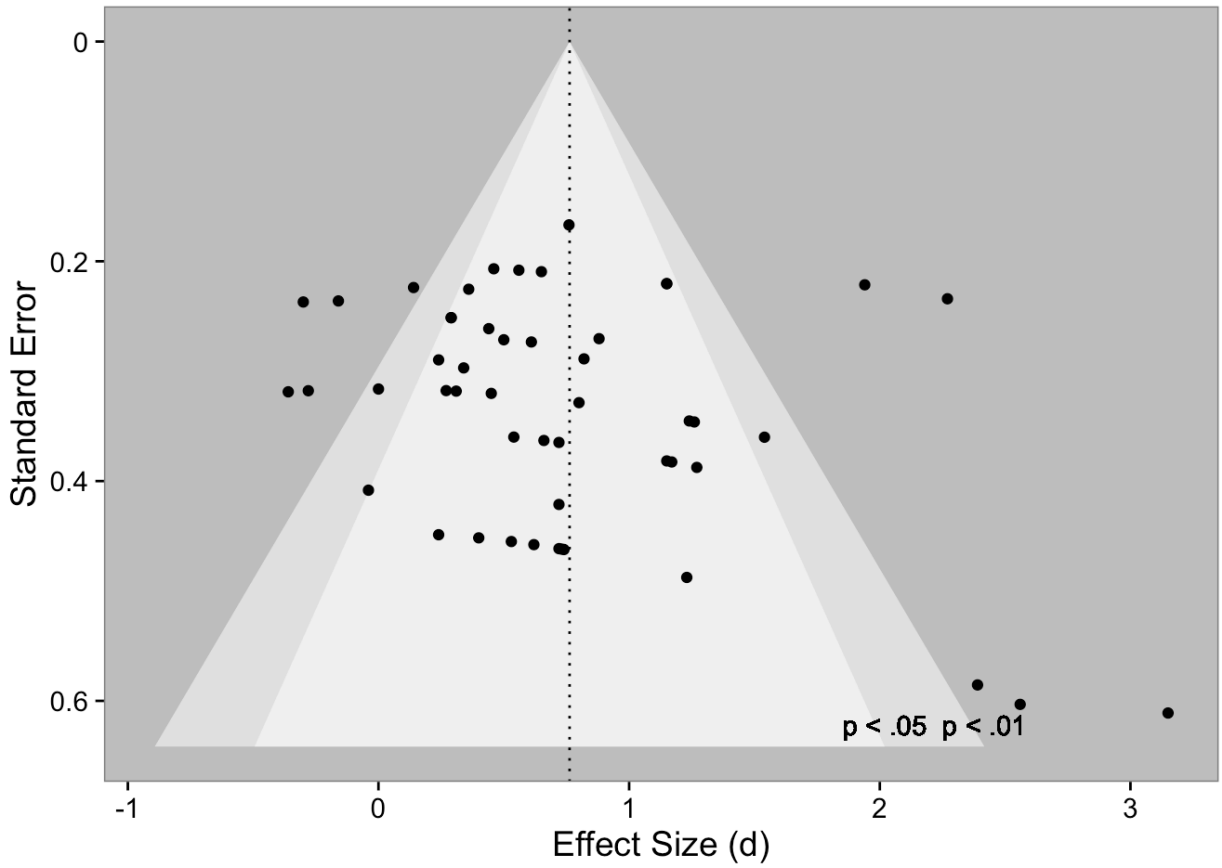

Figure 2. Funnel plot showing the relationship between standard error and effect size for studies of IDS preference, modified from http://metalab.stanford.edu. Individual dots represent studies. Larger and smaller funnel boundaries show 99% and 95% thresholds, respectively. Dotted line shows the mean effect size from a random-effects meta analytic model.