A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Molly Lewis[1], Mika Braginsky[1], Sho Tsuji[2], Christina Bergmann[2], Page Piccinini[2], Alejandrina Cristia[2], Michael C. Frank[1]

[1] Department Psychology, Stanford University

[2] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS

Author note

Correspondence concerning this article should be addressed to Molly Lewis, Psychology Department, Stanford University. 450 Serra Mall, Stanford, CA 94305. E-mail: mll@stanford.edu.

Abstract

replicability, etc.

Keywords: replicability, reproducibility, meta-analysis, developmental psychology, language acquisition

Word count: XXXX

A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Introduction

Psychologists hope to build generalizable theories about human behavior—theories that hold true beyond particulars of an individual study. The field has grown concerned as a result in the face of recent high-profile evidence that an effect observed in one study may not be the same in another ("replicability crisis"; Ioannidis, 2005; Nosek, 2012, 2015). Some of this variability is to be expected, however—the question we should instead be asking is, do the data provide support for the theory, even if they are noisy? Furthermore, to build parsimonious theories of human behavior, we should seek to explain not just individual phenemenon, but entire literatures of research. What is needed, then, is a tool for aggregating noisy data across studies within a phenomenon, as well as a common language for comparing effects across phenomenona.

Meta-analytic methods provide a powerful tool for doing just this. The basic unit of meta-analysis—the effect size—provides an estimate of the *size* of an effect, as well as a measure of uncertainty around this point estimate. With such a continuous measure of success, we can apply the same reasoning we use to aggregate noisy measurements over participants in a single study: By assuming each *study*, rather than participant, is sampled from a population, we can appeal to the classical statistical framework to combine estimates of the effect size for a given phenomenon.

This quantitative approach provides a rich tool kit for synthesizing across literatures. By describing different phenomena using the same unit of measurement, we are able to compare effects in different domains. Rather than simply concluding that two effects are both "real," we can ask more fine-grained questions: Is effect X bigger than effect Y? Does a moderator influence effect X in the same way as effect Y? This type of continuous analysis supports building quantitative models, and specifying theories that are more precise and constraining.

In addition to these theoretical motivations, there are practical reasons for conducting

a quantitative synthesis. When planning an experiment, an estimate of the size of an effect on the basis of prior literature can inform the sample size needed to achieve a desired level of power. Meta-analytic estimates of effect sizes can also aid in design choices: If a certain paradigm tends to have overall larger effect sizes than another, the strategic researcher might select this paradigm in order to maximize the power of a study.

In practice, however, the feasability of this meta-analytic approach relies on the field's commitment to practices that facilitate cumulative science. These practices apply to all stages of the research process. At the stage of experimental planning, researchers must pre-specify analytical descision to limit "researcher" degrees of freedom (Simmons, 2011; Simonsohn, 2014a, 2014b, 2014c). At the stage of completion, researchers should share a result regardless of its significance (Rosenthal, 1979; Fanelli 2012). And, at the stage of sharing, researchers must provide enough information about the method for another lab to conduct a close replication. Critically,r eports must also contain complete descriptions of both data and analytical decisions so that effect sizes can be calculcated for the purposes of meta-analysis,

In the present paper, we use meta-analytic methods to provide a quantitative synthesis of an entire field of psychological research: language acquisition. We think this field is a particularly informative case study. It may be particularly vulnerable to false findings because running children is expensive (Ioanndis, 2005), and thus:

- sample sizes are small
- replications difficult and rare
- Recent attention about practices in developmental research Peterson (2016)

We have two goals:

- Describe the state of the field in terms of its participation in practices that are prerequisites to cumulative science, and ultimately, a theoretical synthesis
- Provide a preliminary theoretical synthesis of the field

Towards this end, we introduce Metalab.

Method

We calculated estimates of effect sizes for 11 different phenomenena in language acquisition. We selected these phenomena in order to describe development at many different levels of the language hierarchy, from the acquistion of prosody and phonemic contrasts, to gaze following in linguistic interaction. This wide range of phenomena allowed us to compare the course of development across different domains, as well as explore questions about the interactive nature of language acquisition.

Estimates of effect size were based on journal reports of experimental data. In total, our sample includes estimates from 258 papers, 938 different conditions and 11,628 participants.

The process for selecting papers from the literature differed by domain, with some individual meta-analyses using more systematic approaches than others. [Simulations here?]

Replicability of the field

Effect size can vary between studies for reasons unrelated to a theoretical construct. One reason for this variability is the precision of the effect size, which we can model based on the sample size of the study. A remaining source variability, however, are biases introduced directly by the experimenter, via publication bias (Fanelli, 2010; Rosenthal, 1979; Rothstein, Sutton, & Borenstein, 2006), analytical flexibility (Simmons, Nelson, & Simonsohn, 2011), reporting errors, or even fraud. These biases are much more difficult to model, and may therefore lead to large but unknown errors in estimates of the effect size. If these types of practices are present in the literature, estimates of effect size may be poor estimates of the true underlying effect size, making it difficult to make theoretical progress. Below we present analyses examining whether signatures of publication bias and analytical flexibility are present in the language acquistion literature. We find little evivdence of these biases.

Level	Phenomenon	Description	N papers (conditions)
Prosody	IDS preference (Dunst, Gorman, & Hamby, 2012)	Looking times as a function of whether infant-directed vs. adult-directed speech is presented as stimulation.	16 (50)
Sounds	Phonotactic learning (Cristia, in prep)	Infants' ability to learn phonotactic generalizations from a short exposure.	15 (47)
	Vowel discrimination (native) (Tsuji & Cristia, 2014)	Discrimination of native-language vowels, including results from a variety of methods.	40 (167)
	Vowel discrimination (non-native) (Tsuji & Cristia, 2014)	Discrimination of non-native vowels, including results from a variety of methods.	21 (72)
Phonotactics	Statistical sound learning (Cristia, in prep)	Infants' ability to learn sound categories from their acoustic distribution.	11 (40)
Proto-words	Word segmentation (Bergmann & Cristia, 2015)	Recognition of familiarized words from running, natural speech using behavioral methods.	67 (295)
Words	Mutual exclusivity (Lewis & Frank, in prep)	Mapping of novel words reflecting children's inference that novel words tend to refer to novel objects.	20 (60)
	Concept-label advantage (Lewis & Long, unpublished)	Infants' categorization judgments in the presence and absence of labels.	16 (100)
	Online word recognition (Frank, Lewis, & MacDonald, 2016)	Online word recognition of familiar words using two-alternative forced choice preferential looking.	12 (32)
Communication	Gaze following (Frank, Lewis, & MacDonald, 2016)	Gaze following using standard multi-alternative forced-choice paradigms.	15 (45)
	Pointing and vocabulary (Colonnesi et al., 2010)	Longitudinal correlations between declarative pointing and later vocabulary.	25 (30)

p-curves

(Simonsohn, Nelson, & Simmons, 2014a, 2014b; Simonsohn, Simmons, & Nelson, 2015) Across studies we should expect some variability in effect size due to sampling error alone. But this variability in effect size should be *systematic*: There should be less variability around the mean for more precise studies, as measured by sample size. The presence of variability in effect sizes that is not accounted for sample size may suggest publication bias in a literature.

Phenomenon	$\bar{d}$	power	p-curve skew	funnel skew	fail-safe-N
IDS preference	0.71 [0.53, 0.89]				
Phonotactic learning	0.04 [-0.09, 0.16]				
Vowel discrimintation (native)	0.6 [0.5, 0.71]				
Vowel discrimination (non-native)	0.66 [0.42, 0.9]				
Statistical sound learning	-0.14 [-0.27, -0.02]				
Word segmentation	0.19 [0.14, 0.24]				
Mutual exclusivity	1 [0.68, 1.33]				
Concept-label advantage	0.4 [0.29, 0.51]				
Gaze following	0.84 [0.26, 1.42]				
Pointing and vocabulary	0.41 [0.32, 0.49]				

bias introduced by meta-analysis in selection (second-order selection bias)

Funnel

- Egger's regression test

Orwin

Power

Ioannidis and Trikalinos (2007)

TABLE WITH: Eggers regression, p-curve (stouffer), regular power, p-power, orwin

Theoretical Synthesis

OUTLINE

Statistical Approach

METAMETAPLOT

Discussion

Author Contributions.

Acknowledgments.

References. Bergmann, C., & Cristia, A. (2015). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*.

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

Fanelli, D. (2010). Positive Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*(4), e10068–10.

Frank, M. C., Lewis, M. L., & MacDonald, K. (in press). A performance model for early word learning. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Retrieved from http://langcog.stanford.edu/papers_new/frank-2016-underrev.pdf

Lewis, M., & Frank, M. C. (in prep). Multiple routes to disambituation.

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*(0), 1–10.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons.

Simmons, 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, Nelson, L. D., & Simmons, J. P. (2014a). p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*, *9*(6), 666–681.

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

Simonsohn, Simmons, J. P., & Nelson, L. D. (2015). Better p-curves. *Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson (Forthcoming),"Better P-Curves," Journal of Experimental Psychology: General*.

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