Commentary/Yorkoni: The generalizability crisis

Improving the generalizability of infant psychological research: The ManyBabies model

Ingmar Visser1, Christina Bergmann2, Krista Byers-Heinlein3, Rodrigo Dal Ben4, Wlodzislaw Duch5, Samuel Forbes6, Laura Franchin7, Michael C. Frank8, Alessandra Geraci9, J. Kiley Hamlin10, Zsuzsa Kaldy11, Louisa Kukel12, Catherine Laverty13, Casey Lew-Williams14, Victoria Mateu15, Julien Mayor16, David Moreau17, Iris Nomikou18, Tobias Schuwerk19, Elizabeth A. Simpson20, Leher Singh21, Melanie Soderstrom22, Jessica Sullivan23, Marion I. van den Heuvel24, Gert Westermann25, Yuki Yamada26, Lorijn Zaadnoordijk27 and Martin Zettersten1

1Department of Psychology, University of Amsterdam, Amsterdam, 1018 WB, The Netherlands; 2Language and Development Department, Max Planck Institute for Psycholinguistics, 6525 XD Nijmegen, The Netherlands; 3Concordia Infant Research Laboratory, Concordia University, Montreal QC H4B 1R6, Canada; 4Nicolaus Copernicus University, 87-100 Torun, Poland; 5University of East Anglia, Norwich NR4 7TJ, UK; 6Department of Psychology and Cognitive Science, University of Trento, 38068 Rovereto, Italy; 7Stanford University, Stanford, CA 94301 USA; 8UBC Center for Infant Cognition, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; 9UMass Boston, Baby Lab, Department of Psychology, University of Massachusetts Boston, Boston, MA 02125-3393, USA; 10Neurocognitive Developmental Psychology, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Germany; 11School of Psychology, University of Birmingham, B15 2TT Birmingham, UK; 12Princeton Baby Lab, Princeton University, Princeton, NJ 08540, USA; 13UCLA Department of Spanish and Portuguese, University of California, Los Angeles, Los Angeles, CA 90095-1532, USA; 14Department of Psychology, University of Oslo, 0373 Oslo, Norway; 15Brain Dynamics Lab, University of Auckland, Auckland 1010, New Zealand; 16Department of Psychology, University of Portsmouth, Portsmouth, UK; 17Department of Psychology, Ludwig Maximilians-Universität München, 80802 Munich, Germany; 18Social Cognition Laboratory, University of Miami, Coral Gables, FL 33124, USA; 19Department of Psychology, National University of Singapore, Singapore 119077; 20Baby Language Lab, University of Manitoba, Winnipeg, MB R3T 2N2, Canada; 21Developing Minds Center, Skidmore College, Saratoga Springs, NY 12866, USA; 22Department of Cognitive Neuropsychology, Tilburg University, 5037 AB Tilburg, The Netherlands; 23Department of Psychology, Lancaster University, Lancaster LA1 4YW, UK; 24Kyushu University, Fukuoka, Japan and Trinity College Dublin, Dublin, Ireland

i.visser@uva.nl
christina.bergmann@mpi.nl
k.byers@concordia.ca
dalbenwork@gmail.com
i.w Duch@uvm.pl
samuel.forbes@uea.ac.uk
laura.franchin@unitn.it
mcfrank@stanford.edu
alessandra.geraci@unitn.it
kiley.hamlin@psych.ubc.ca
zsuzsa.kaldy@umb.edu
louisa.kukel@fau.de
CML704@student.bham.ac.uk
caseylw@princeton.edu

Downloaded from https://www.cambridge.org/core. University of Massachusetts - Boston (Healey Library), on 10 Feb 2022 at 14:53:02, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0140525X20001685
Yarkoni raises concerns about widespread practices in the psychological sciences – ranging from standard statistical practices to narrow experimental designs – which hinder generalizability, theory-building, and ultimately, explanatory power. Infant research in particular faces a range of problems, including difficulties recruiting participants (often resulting in small samples), the unique challenges of designing experiments that hold infants’ attention, limited numbers of observations per participant, and infants’ rapid developmental changes (Bergmann et al., 2018; Frank et al., 2017; Oakes, 2017).

Many Babies is a large-scale, multilab collaborative project that currently spans 47 countries and over 200 institutions (https://manybabies.github.io). The project provides a constructive, best-practice, grass-roots approach for addressing issues of replicability and generalizability in infant research and employs a model also utilized by other large-scale, multisite collaborations (e.g., ManyPrimates, 2019; Moshontz et al., 2018). Thus far, Many Babies has focused its efforts on replicating fundamental findings in infant cognition that underpin our understanding of early cognitive development.

Features and benefits of the Many Babies approach in addressing the issues Yarkoni identified are (see also Byers-Heinlein et al., 2020; Frank et al., 2017; The Many Babies Consortium, 2020):

1. **Consensus-based study designs to advance theory.** Many Babies projects are focused on evaluating central theories in infant research (e.g., under which circumstances infants show preferences for familiar or novel stimuli in Many Babies5; Hunter & Ames, 1988), and carefully probing the bounds of theoretical constructs by encouraging participation from researchers with diverse perspectives. Many Babies’ collaborative and consensus-building approach disrupts existing hierarchies, making space for dissent and innovation, and for adjudicating between opposing views (e.g., in the case of adversarial collaboration in Many Babies2 addressing Theory of Mind; c.f. Baillargeon, Buehmann, & Southgate, 2018; Cowan et al., 2020; Surian & Geraci, 2012). Simultaneously, it expands collaborative networks to bridge a wide variety of theoretical backgrounds, resulting in designs that clearly identify testable points of disagreement to lay the foundation for further inquiry through experiment and debate.

2. **Conceptual replications.** As noted by Yarkoni, direct replication is not a sensible target for improving reproducibility if there are concerns about weaknesses in paradigms or stimulus sets that could be addressed in a new experiment (e.g., Many Babies4 will remove confounds in a paradigm developed to probe infants’ social evaluations; Hamlin, Wynn, & Bloom, 2007; Scarf, Imuta, Colombo, & Hayne, 2012). Many Babies projects probe the generality of phenomena by prioritizing conceptual over exact replications, bringing together researchers from different theoretical and methodological backgrounds to build experimental designs that best capture the processes being studied.

3. **Diversity in samples and scientists.** By encouraging participation from labs from all over the world and supporting laboratory expenses for scientists who are new to experimental infant research, Many Babies promotes diversity across multiple dimensions: contexts, lab practices, researchers, and participants. Many Babies takes seriously the importance and impact of participant heterogeneity (Henrich, Heine, & Norenzayan, 2010), and creates datasets that are more representative of the population of interest (i.e., “human infants”) compared to single-lab studies, by testing participants with diverse linguistic and sociocultural backgrounds. Exploring the impact of diversity on the generalizability of core findings has become a prominent target in recent projects, e.g., studying infants at home rather than in a highly-controlled lab setting in Many Babies-At-Home, thereby reaching more rural populations; assessing the replicability of initial findings with African infants in Many Babies1A; in Many Babies3 – studying rule-learning – making the stimuli suitable for infants from different linguistic backgrounds. In doing so, Many Babies enables us to strike a better balance between
the precision of estimation/breadth of generalization trade-off cited by Yarkoni.

(4) **Quantifying sources of variation.** Studies following the ManyBabies approach can reveal and explicitly measure sources of variation that are difficult to estimate in single-lab studies, including effects of lab practices and methodological variation. For example, ManyBabies1 (addressing infants’ preferences for infant-directed speech) tested for effects of distinct experimental methods in infant research (e.g., head-turn preference, central fixation, eye-tracking, ManyBabies Consortium, 2020); ManyBabies2 compares online and in-lab data collection. Both projects thereby probe the generalizability of observed phenomena across experimental paradigms. Specifically, variety is built in through diversity of experimental paradigms used to test a research question—a typical benefit of meta-analysis—yet at the same time we retain control over a number of design factors, as in replication efforts. Given the wide-ranging sources of methodological variation, however, there is considerable work remaining to be done on this issue.

(5) **Stimulus generalizability.** Issues related to stimulus informativeness and generalizability (or lack thereof) are discussed by the ManyBabies project teams and wider community throughout the design process, which generates new “best test” stimuli. The focus is on conceptual replications that involve stimulus sets that differ from the original studies, in this way directly addressing the question of stimulus generalizability. The next step here is to systematically vary stimulus sets.

(6) **Transparent research practices.** ManyBabies is committed to transparency at each research stage, and to collective governance that encourages genuine and non-hierarchical debate, defies the research status-quo, and leads to innovation in theoretical, methodological, and analytic design, as Yarkoni suggests. For example, ManyBabies maintains detailed documentation protocols and openly shares all stimuli and data, including many additional descriptive variables. In this way, additional sources of variance and alternative hypotheses can be tested.

Ensuring that verbal and quantitative expressions of our hypotheses are closely aligned is a tall task. The diversity of scientists involved in each ManyBabies project goes a long way toward developing meaningful operationalizations of the specific research questions under examination. At the same time, the diversity of sources, methods, and stimuli addresses (to an extent) many of the questions on generalizability raised by Yarkoni. Even so, much work remains to tackle concerns related to methodological/stimulus variation, generalizability, and participant heterogeneity, to develop best practices in large-scale international collaborations, and to build better theories (Borsboom, van der Maas, Dalege, Kievit, & Haig, 2021). Nevertheless, we look forward to continuing to provide opportunities for learning and growth in the ManyBabies communities, creating the necessary scaffolding for even better research, and, alongside other large collaborative networks, being at the forefront of creating a psychological science that is generalizable and reproducible.

**Financial support.** This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

**Conflict of interest.** None.

---

**References**


---

**An accelerating crisis: Metascience is out-reproducing psychological science**

**Patrick D. Watson**

Minerva Schools at the Keck Graduate Institute, San Francisco, CA 94103, USA.
pwatson@minerva.kgi.edu; https://www.patrickdkw.com/

doi:10.1017/S0140525X21000121, e36

**Abstract**

Scientific claims are selected in part for their ability to survive. Scientists can pursue an r-strategy of broad, easy-to-spread ideas, or a K-strategy of stress-tested, bulletproof statements.