communications psychology

A Nature Portfolio journal

Comment

https://doi.org/10.1038/s44271-026-00477-w

What 5000 babies can tell us about developing minds and how to study them

Brianna T. M. McMillan, Heidi A. Baumgartner, Christina Bergmann, Michael C. Frank, J. Kiley Hamlin, Dora Kampis, Melissa Kline Struhl, Eon-Suk Ko, Jessica E. Kosie, Casey Lew-Williams, Kelsey Lucca, Tobias Schuwerk, Melanie Soderstrom, Ingmar Visser, Francis L. Yuen, Martin Zettersten & Krista Byers-Heinlein

Check for updates

A decade of ManyBabies research, testing thousands of babies across hundreds of labs, has shown that some, but not all findings in infant research replicate well. Collectively, these projects have shown us that our methods carry limitations that larger samples alone cannot resolve. Here we present three lessons that point toward a more reliable, inclusive developmental science.

Research on infant cognitive development seeks to understand how the mind is structured, how it changes with experience, and what this reveals about human psychology more broadly. For much of its history, infant studies have been almost exclusively conducted in individual labs with a relatively small number of infants. The single-lab model yields remarkable insights but produces work that is often small in scale, difficult to replicate, narrow in cultural scope, and hard to generalize across populations.

The ManyBabies Consortium was founded in 2015 to tackle these challenges[1], and has grown to involve over 800 researchers in more than 50 countries across six continents. Like the ManyLabs projects[2], the initial goal was replication at scale. However, babies are not straightforward participants: they have short, unpredictable attention spans, limited behavioral repertoires, and, because of their reliance on caregivers, are difficult to study in controlled experimental settings. By studying thousands of babies across more than 200 labs worldwide, the challenges of infant research have come into sharper focus. In the first decade of ManyBabies research, three lessons have emerged that reshaped how we think about studying infant development: large variability across infants, labs, and contexts turned out to be the norm, not the exception (Lesson 1); larger samples sharpened rather than resolved our measurement challenges (Lesson 2); and methodological choices are cultural choices (Lesson 3).

Lesson 1: variability is data

In a typical infant experiment, the infant sits on the caregiver's lap in a quiet room, and a researcher presents them with stimuli, such as audio clips, images on a screen, or a researcher engaging the infant across a table. Unlike adult participants, infants cannot follow instructions or report thoughts and feelings, so researchers typically rely on observable behaviors, such as whether infants look longer at one stimulus than another, reach toward it, or turn away. These subtle, fleeting behavioral responses are the primary data of behavioral infant research.

In our first collaborative project (MB1: Infant-Directed Speech Preference[3]), we scaled up to 67 labs to investigate what many researchers consider one of the most robust phenomena in developmental science: infants' preference for infant-directed speech–the high-pitched, melodic way adults in most–but not all–cultures naturally talk to babies. The overall finding was that infants prefer infant-directed speech over adult-directed speech. But responses varied enormously within and across labs: some labs found strong preferences for infant-directed speech; others found weak or negligible effects; a handful found preferences for adult-directed speech; individual responses varied widely (Fig. 1).

Our field has long treated this kind of variability as "noise", something to be minimized through careful experimental control, larger samples, or better standardization. Sometimes that framing is right as not all variability is theoretically meaningful (e.g., an infant distracted by their sock is not telling us much about speech perception), and some portion of it reflects simple measurement error. But infant behavior is also shaped by factors we typically do not measure, for example, temperament, prior experience, testing context, and caregiver behavior. Whether any given variability is nuisance or signal is an often unanswered empirical question.

ManyBabies addresses this by balancing experimental control with systematic documentation of contextual variation. Across projects, lab protocols instruct sites to standardize core procedural elements while recording a wide range of additional variables drawn from both the empirical literature and the accumulated intuitions of researchers who have spent years conducting infant studies (e.g., testing room features, experimenter characteristics). The scale of ManyBabies, combined with open, depersonalized data sharing, enables pooling individual-level data and systematically testing which factors drive variability. In MB1, for instance, some of the variability we observed related to age, infant language background, and experimental method. Collecting this rich contextual information lets us systematically examine drivers of variability, and in doing so, better understand how development unfolds across individuals and contexts.

Lesson 2: scaling up *doesn't* solve all measurement challenges

We initially thought that many of the measurement limitations in infant research–including underpowered designs and idiosyncratic lab effects–might be largely due to small sample sizes, a problem that is solvable by studying more babies across more labs. In practice, larger samples, while solving a critical part of the problem, often brought measurement challenges into sharper focus rather than resolving them. We organize what we have learned around three properties of measurement: the robustness of our effects, the reliability of our measures, and the validity of our interpretations.

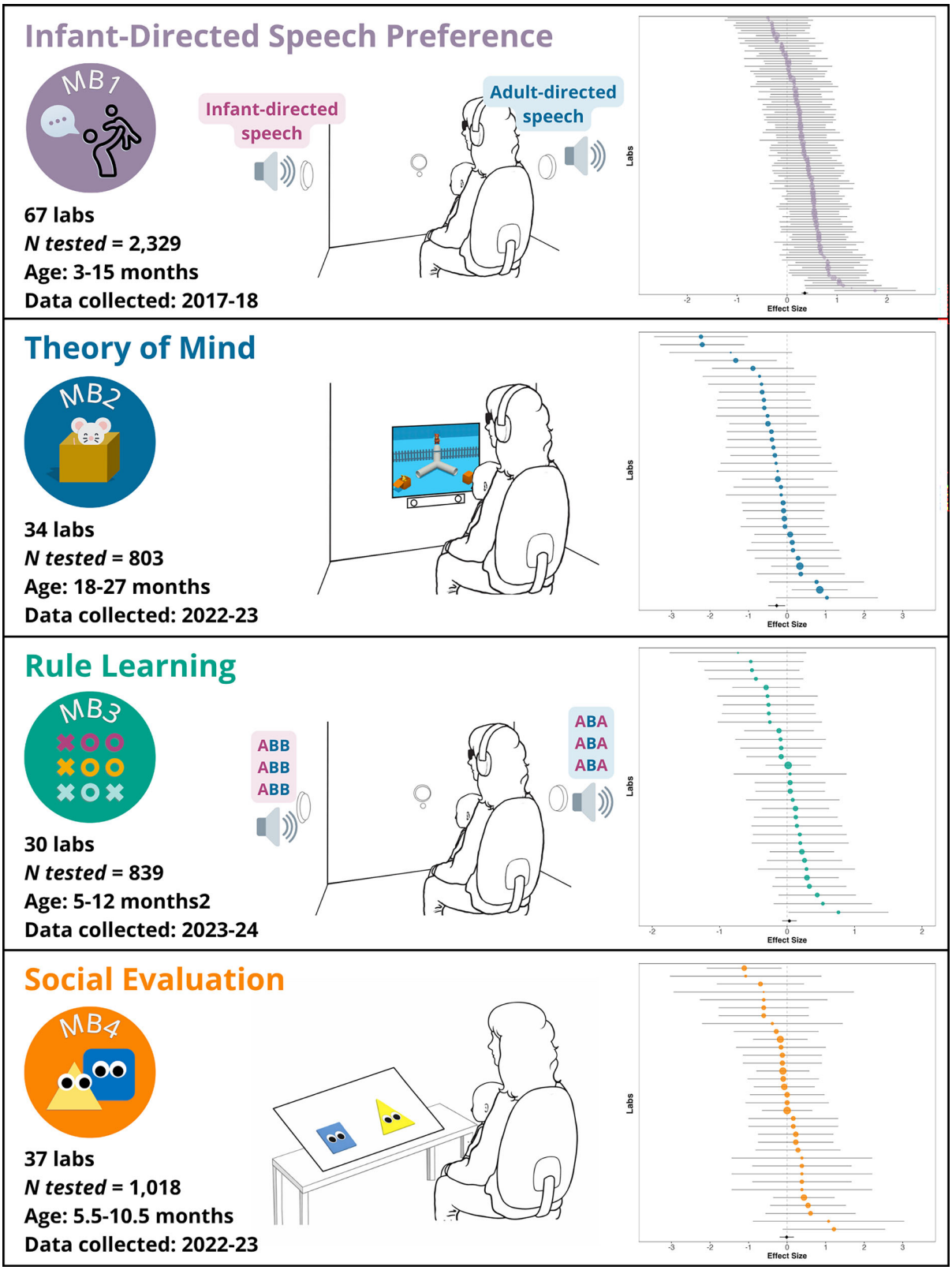
Infant-Directed Speech Preference
MB1
Infant-directed speech
Adult-directed speech
67 labs
N tested = 2,329
Age: 3-15 months
Data collected: 2017-18
Labs
Effect Size
Theory of Mind
MB2
34 labs
N tested = 803
Age: 18-27 months
Data collected: 2022-23
Labs
Effect Size
Rule Learning
MB3
ABB
ABB
ABB
ABA
ABA
ABA
30 labs
N tested = 839
Age: 5-12 months2
Data collected: 2023-24
Labs
Effect Size
Social Evaluation
MB4
37 labs
N tested = 1,018
Age: 5.5-10.5 months
Data collected: 2022-23
Labs
Effect Size

Fig. 1 | Variability in effect size within and across labs in four ManyBabies studies. This figure illustrates the different methods used in the first four ManyBabies studies (MB1, MB2, MB3, and MB4), including the number of participating labs, sample size, age range of participants, and data collection period. The forest plots show the meta-analytic results for each focal construct. Error bars represent 95% confidence intervals around each lab's standardized effect size. Method drawings are adapted from Gervain & Werker, 2013[10], Figure 4, CC-BY-SA-3.0.

Robustness. The most straightforward benefit of scale is statistical power: larger samples make it possible to detect group-level effects with greater confidence, and to determine whether a finding holds up across different labs and populations. In this sense, ManyBabies has helped. The infant-directed speech preference in MB1, for example, was confirmed across dozens of labs to be a genuinely robust phenomenon. But scale also revealed the fragility of some effects. As shown in Fig. 1, several of our subsequent studies (MB2, MB3, MB4) yielded results that were null or even the opposite of what was predicted. In MB4, for instance, all 37 labs used a single method that had previously demonstrated infants' preference for prosocial over antisocial agents[4]. However, when results did not support the hypothesis, it was difficult to disentangle whether this reflected the phenomenon we were trying to study, the specific paradigms we chose, or other factors, such as some unspecified impact of the COVID-19 pandemic.

Reliability. Researchers have long suspected that many infant measures have limited consistency from one measurement to the next, resulting in low test/retest reliability. Larger samples increase statistical power, making it easier to detect effects in groups of infants, but this does not improve the reliability of measurement for any individual infant. A task that is powerful at the group level may tell us little about whether the same infant would respond similarly from one occasion to the next, or whether infants who show stronger responses in one task show stronger responses in another.

This matters because many of the questions we want to ask (e.g., whether early behaviors predict either later development or responses in other tasks) require precise and reliable individual-level measurement. For example, one spin-off study of the MB1 project asked whether infant-directed speech preference in early infancy predicted later vocabulary size[5]. There was no evidence for such a relationship, but this result is difficult to interpret given the low reliability of the preference measure itself[6]. Developing paradigms that yield more consistent individual-level data is a critical next step.

Validity. The methods we use to study infant cognition often require us to make inferences from behavior. When an infant looks longer at one stimulus than another, what does that actually mean? Surprise? Interest? Confusion? Preference? These interpretations carry different theoretical implications, and our tasks rarely let us distinguish between them. ManyBabies studies are tackling this in two ways. One approach is to improve our fundamental understanding of infants' behaviors themselves. For example, MB5 investigates the well-documented but poorly understood shift from familiarity to novelty preference in infant looking behavior, attempting to identify factors that explain how a given infant in a given study allocates their attention. A second approach is to triangulate across behavioral, neural, and physiological measures. For example, the ManyBabies 3 rule-learning project tests infants' sensitivity to abstract patterns using converging behavioral (looking time), neural (fNIRS), and physiological (pupillometry) measures. When different measures converge, our confidence in construct validity grows; when they diverge, we learn something important about what our tasks are actually capturing[7]. Systematically identifying the factors that shape infant behavior and triangulating across multiple measures are particularly critical for a field whose participants cannot report on their own cognition.

In sum, scaling our sample sizes is helpful but insufficient. Resolving measurement challenges might require improving current paradigms or even developing new ones.

Lesson 3: scaling up can make cultural assumptions visible

Infant development is shaped by cultural context, and multi-site research surfaces this in ways that are easy to overlook in single-lab studies. In planning MB1, we chose to use a single set of North American English stimuli, prioritizing standardization over ecological validity. The data confirmed that this choice carried cultural assumptions: infants who were growing up in North American English contexts showed stronger infant-directed speech preferences than infants learning other languages[3]. Without a basis for comparison, a single-lab finding cannot distinguish universality from a context-specific effect.

The tension between standardization and customization generated spin-off projects designed to disentangle universal from experience-dependent effects, including testing infant-directed speech preference in infants' native languages, examining how bilingual language exposure shapes the preference, and extending the work to African populations absent from MB1, including an ongoing study by Tsui and colleagues. Even describing participants consistently across cultures required its own dedicated effort: basic demographic categories like socioeconomic status, bilingualism, and race that seem straightforward in a North American context do not translate cleanly across cultures, a challenge addressed in a dedicated methodological paper[8]. We are also broadening the questions we ask, for example, examining early childhood media use across cultural contexts. ManyBabies' reach into the plurality of human cultural contexts remains limited[9], but the multi-site structure has made the cultural specificity of our methods visible in a way that single-lab research rarely does.

Conclusion

ManyBabies began as an effort to establish whether key findings in infant research could be replicated across labs. Over a decade on, that question remains important, but our experience has pushed us toward a broader set of concerns. Scaling up has helped: larger samples and open data sharing have enabled more reliable detection of effects and investigation of sources of variability that single-lab studies cannot. But scale alone cannot resolve the challenges we have encountered. What is also needed is a clearer understanding of how to amplify genuine signals and diminish noise in infant data, how to measure meaningful sources of variability at the individual level, and how to ensure that our questions, stimuli, and procedures are meaningful across the full range of human cultural contexts.

The collaborative structure of ManyBabies has also been essential for generating the data and the cross-lab conversations that made these realizations possible in the first place. Working across hundreds of labs and multiple continents has united researchers around shared problems in a way that single-lab work rarely does. The lessons we have drawn here about variability, measurement, and cultural context are ones that the field is now better positioned to address collectively, and we hope they resonate beyond large-scale collaborations. Single-lab studies also benefit from open data

Table 1 | Applying open science practices to the three core lessons from ManyBabies studies: Recommendations for future research

Lesson	Open data and materials	Documentation of data collection	Collaborative norms
Lesson 1: variability is data	Share de-identified, individual-level data, standardized metadata, and demographics to enable reanalysis and pooling	Document testing contexts systematically through walk-through videos and structured session logs that capture infant state, caregiver behavior, and room features	Coordinate diverse sampling, preregister shared moderators, and co-design measures
Lesson 2: scaling up doesn't solve all measurement challenges	Publish stimuli, task code, preprocessing and analysis scripts, and preregistrations for reanalysis and robustness checks	Keep trial-level logs, timestamps, coder notes, calibration/failure reports, and raw recordings when possible. Where feasible, collect repeated measures to assess test-retest reliability	Collaborate to develop new, high-reliability protocols, run cross-site pilots, use shared protocols and centralized quality control, implement inter-rater reliability checks, and triangulate across behavioral, neural, and physiological measures
Lesson 3: scaling up can make cultural assumptions visible	Share stimulus files, translations, and detailed participant descriptors to enable cultural reanalysis	Document culturally relevant participant information (e.g, stimuli familiarity, lab location, language experience) to identify cultural moderators rather than assume universality	Collaborate with local partners at all stages of research, adapt stimuli when appropriate, and ensure equitable credit

practices, richer documentation of data collection procedures, and collaborative norms that projects like ManyBabies have helped to establish (Table 1).

Studying so many babies has not always yielded the clear answers we hoped for. Null results, ambiguous findings, and hard questions about how to design our studies and interpret our data have been a recurring part of the ManyBabies experience. But confronting these problems at scale and together has been more productive than addressing them quietly in individual labs. Our field now has a clearer view of where infant behavioral research stands, what current methods can and cannot deliver, and what it will take to make progress. That, too, is something that many babies have taught us.

Brianna T. M. McMillan [1] ✉, Heidi A. Baumgartner [2] ✉, Christina Bergmann [3,4], Michael C. Frank [2], J. Kiley Hamlin [5], Dora Kampis [6], Melissa Kline Struhl [7], Eon-Suk Ko [8], Jessica E. Kosie [9], Casey Lew-Williams [10], Kelsey Lucca [11], Tobias Schuwerk [12], Melanie Soderstrom [13], Ingmar Visser [14], Francis L. Yuen [5], Martin Zettersten [15] & Krista Byers-Heinlein [16]

[1]Department of Psychology, Smith College, Northampton, MA, USA. [2]Department of Psychology, Stanford University, Stanford, CA, USA. [3]Osnabrück University of Applied Sciences, Osnabrück, Germany. [4]Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands. [5]Department of Psychology, University of British Columbia, Vancouver, BC, Canada. [6]Department of Psychology, University of Copenhagen, Copenhagen, Denmark. [7]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [8]Department of English Language and Literature, Chosun University, Gwangju, South Korea. [9]School of Social and Behavioral Sciences, Arizona State University, Phoenix, AZ, USA. [10]Department of Psychology, Princeton University, Princeton, NJ, USA. [11]Department of Psychology, Arizona State University, Tempe, AZ, USA. [12]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany. [13]Department of Psychology, University of Manitoba, Winnipeg, MB, Canada. [14]Department of Psychology, University of Amsterdam, Amsterdam, Netherlands. [15]Department of Cognitive Science, University of California San Diego, San Diego, CA, USA. [16]Department of Psychology, Concordia University, Montréal, QC, Canada.
✉e-mail: bmcmillan@smith.edu; heidib@stanford.edu

Received: 20 April 2026; Accepted: 20 May 2026;
Published online: 09 June 2026

References

1. Frank, M. C. et al. A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* **22**, 421–435 (2017).
2. Klein, R. A. et al. Investigating variation in replicability. *Soc. Psychol.* **45**, 142–152 (2014).
3. ManyBabies Consortium. Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* **3**, 24–52 (2020).
4. Lucca, K. et al. Infants' social evaluation of helpers and hinderers: a large-scale, multi-lab, coordinated replication study. *Dev. Sci.* **28**, e13581 (2025).
5. Soderstrom, M. et al. Testing the relationship between preferences for infant-directed speech and vocabulary development: a multi-lab study. *J. Child Lang.* **52**, 984–1009 (2025).
6. Schreiner, M. S. et al. Limited evidence of test-retest reliability in infant-directed speech preference in a large preregistered infant experiment. *Dev. Sci.* **27**, e13551 (2024).
7. Havron, N. Why not both? Using multiple measures to improve reliability in infant studies. *Infant Child Dev.* **31**, e2336 (2022).
8. Singh, L. et al. A unified approach to demographic data collection for research with young children across diverse cultures. *Dev. Psychol.* **60**, 211–227 (2024).
9. Jia, F. The tension between big team science and colonial power dynamics. *Nat. Hum. Behav.* **10**, 440–442 (2026).
10. Gervain, J. & Werker, J. F. Prosody cues word order in 7-month-old bilingual infants. *Nat. Commun.* **4**, 1490 (2013).

Acknowledgements

H.A.B., J.K.H, M.S., F.L.Y., and K.-B.H. disclose support for the research of this work from the Social Sciences and Humanities Research Council of Canada Partnership Grant (895-2025-1008). J.E.K. and M.Z. disclose support for the research of this work from the Einstein Foundation Early Career Award. E.-S.K. discloses support for the research of this work from the National Research Foundation of Korea (NRF-2025S1A6B5A02004207). T.S. discloses support for the research of this work from the Deutsche Forschungsgemeinschaft (DFG SCHU 3060/2-2). C.L.W. discloses support from the Data-Driven Social Sciences Grant and the Global Collaborative Network Grant, Princeton University. B.T.M.M., C.B., M.C.F., D.K., M.K.S., K.L., and I.V. declare no relevant funding.

Author contributions

Conceptualization: B.T.M.M., H.A.B., C.B., M.C.F., J.K.H., D.K., M.K.S., E.-S.K., J.E.K., C.L.-W., K.L., T.S., M.S., I.V., F.L.Y., M.Z., and K.B.-H. Project administration: B.T.M.M. and H.A.B. Visualization: H.A.B. and M.Z. Writing - original draft: B.T.M.M., H.A.B., C.B., M.C.F., J.K.H., D.K., M.K.S., E.-S.K., J.E.K., C.L.-W., K.L., T.S., M.S., I.V., F.L.Y., M.Z., and K.B.-H. Writing - review & editing: B.T.M.M., H.A.B., C.B., M.C.F., J.K.H., D.K., M.K.S., E.-S.K., J.E.K., C.L.-W., K.L., T.S., M.S., I.V., F.L.Y., M.Z., and K.B.-H.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44271-026-00477-w.

Correspondence and requests for materials should be addressed to Brianna T. M. McMillan or Heidi A. Baumgartner.

Peer review information The manuscript was considered suitable for publication without further review at *Communications Psychology*. Primary Handling Editor: Jennifer Bellingtier. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026