Received: 26 April 2021 | Revised: 1 December 2021 | Accepted: 12 December 2021

DOI: 10.1002/icd.2295

COMMENTARY

WILEY

Why most psychological research findings are not even wrong

Anne M. Scheel

Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

Correspondence
Anne M. Scheel, Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands.
Email: anne-scheel@gmx.de

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: Vidi grant 452-17-013

Handling Editor: Moin Syed

Abstract

Psychology's replication crisis is typically conceptualized as the insight that the published literature contains a worrying amount of unreplicable, false-positive findings. At the same time, meta-scientific attempts to assess the crisis in more detail have reported substantial difficulties in identifying unambiguous definitions of the scientific claims in published articles and determining how they are connected to the presented evidence. I argue that most claims in the literature are so critically underspecified that attempts to empirically evaluate them are doomed to failure—they are *not even wrong*. Meta-scientists should beware of the flawed assumption that the psychological literature is a collection of well-defined claims. To move beyond the crisis, psychologists must reconsider and rebuild the conceptual basis of their hypotheses before trying to test them.

KEYWORDS
falsification, hypothesis testing, replication crisis, reproducibility, scientific inference

The replication crisis in psychology can be summarized as the field's acknowledgement that a substantial portion of its published findings may be false (Pashler & Wagenmakers, 2012; Shrout & Rodgers, 2018). This statement sounds straightforward enough, and it comes with evidence: In a recent review, Nosek et al. (2022) summarized the outcomes of 307 replications conducted across 77 replication projects in psychology and adjacent fields and found

This work was funded by Vidi grant 452-17-013 from the Dutch Research Council (NWO). As the sole author, A. M. Scheel was responsible for conceptualizing, drafting, and revising this commentary. The author thanks Julia Rohrer and Ruben Arslan for helpful discussions during the preparation of the manuscript.

that only 64% obtained a significant result in the same direction as the original study. Even before large-scale efforts such as the Reproducibility Project: Psychology (Open Science Collaboration, 2015) demonstrated less-than-ideal replicability, simulations and empirical assessments of the literature suggested that an alarming number of published results could be false positives (Greenwald, 1975; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995).

However, the seemingly straightforward statement that a substantial portion of published findings in psychology may be false rests on a strong assumption. The statement assumes that the psychological literature is a collection of individual, well-defined claims—defined well enough, anyway, to be either true or false. The assumption then allows us to quantify the falseness in the literature by examining the statistical results associated with each scientific claim and by attempting to replicate these results with new data. In the following, I will argue that this assumption is wrong and that the underlying problem of ill-defined scientific claims hampers meta-research and current attempts to build a more reliable scientific literature.

1 | MOST SCIENTIFIC CLAIMS IN PSYCHOLOGY ARE ILL-DEFINED

Early in my PhD, I planned a comprehensive assessment of hypotheses and results in published Registered Reports (Chambers, 2013) and preregistered studies, with the idea of creating a database of research findings that should be relatively unaffected by 'questionable research practices' and publication bias. Although the task seemed almost trivial at first, we had to abandon the project after the pilot phase: For most articles in the pilot sample, we were unable to establish exactly which test results informed which hypotheses and how they affected the authors' conclusions. Worse, in many cases, not even the hypotheses themselves were unequivocally stated. Sometimes, they were phrased in such vague terms that it was unclear how they could be operationalized and tested, other times, multiple different, incongruent statements were used to refer to them throughout the manuscript. The articles we analysed all claimed to provide evidence for (or against) *something*, but too often, it was impossible to determine what exactly that something was and what data would support or contradict it—the scientific claims were ill-defined.

A *scientific claim* or *finding*[1] is an inference about the natural world, expressed as an existence statement such as '3-month-olds prefer infant-directed speech over adult-directed speech'. A claim's scope goes beyond a specific set of data points, but it may be fairly close to observable data or further removed from it. Although the project described above focused on hypothesis tests, I remain agnostic about the type of inference that gives rise to a claim—researchers may inductively infer a general pattern from observed data, abductively infer the best explanation for observed data, or hypothetico-deductively infer that a tested hypothesis makes successful predictions (e.g., Fidler, Singleton Thorn, Barnett, Kambouris, & Kruger, 2018). Put differently, a claim can just as well be the product of a serendipitous insight or an open-ended investigation as an a priori prediction that is empirically tested.

As a first basic requirement, scientific claims should be stated unambiguously. But recent meta-scientific studies similar to my own abandoned project show that many published articles fail to clear this low bar. In one such study, Farrar et al. (2020) attempted to survey scientific claims and their statistical evidence in the literature on animal physical cognition. To their surprise, identifying the claims and categorizing them as positive, negative, or inconclusive was so difficult that even a second coding round yielded only 68% inter-rater agreement. In another project, Edelsbrunner and Thurn (2020) analysed the (mis-)interpretation of nonsignificant *p*-values in educational research. They reported similar difficulties in identifying the verbal hypotheses that they encountered, noting that those were often vague statements that encompassed several possible substatements (e.g., six substatements for one hypothesis).

As a second requirement, scientific claims should be accompanied by a definition of their connection to the empirical evidence. A claim cannot stand by itself. Bridging the gap between claim and raw data requires a host of specifications about sampling strategies, measurement instruments and circumstances of the data collection,

analytical decisions, and other auxiliary assumptions. Without specifying this *methodological decision procedure* (Uygun Tunç, Tunç, & Lakens, 2021), we cannot know which observations would constitute evidence for or against the claim—as a consequence, any statistical results are uninterpretable.

The scarcity of such empirical specifications will be painfully familiar to researchers who have tried to independently replicate a published study. It has also proven to be a problem for studying the reproducibility of published meta-analyses. Maassen, van Assen, Nuijten, Olsson-Collentine, and Wicherts (2020) tried to reproduce 33 meta-analyses by retrieving and recomputing the 500 primary effect sizes reported in them. They failed to do so in almost half the cases (224 primary effects) due to 'incomplete or missing information on how effect sizes from primary studies were selected and computed' (p. 1). An earlier attempt by Lakens et al. (2017) to reproduce 20 meta-analyses was ultimately abandoned for the same reason. These examples show that the scientific claims investigated in meta-analyses and primary studies are often too fuzzy to be readily identified and to determine which statistical result(s) inform them, making it difficult (or impossible) to evaluate the evidence for them.

2 | EMPIRICAL UNDERSPECIFICATION HINDERS CONSENSUS ON NEWBORN IMITATION

To take one example, what is the evidence for the claim that humans are born with the ability to imitate others? In 1977, Meltzoff and Moore published the seminal finding that 'human neonates can equate their own unseen behaviors with gestures they see others perform' (p. 75). In two studies on 6 and 12 newborns, respectively, they reported that infants appeared to stick out their tongues, open their mouths, purse their lips, or sequentially move their fingers more often when the experimenter was modelling these gestures than would be expected by chance. The then-typical absence of test and summary statistics (not to mention open data) does not allow us to verify the statistical inferences drawn from the eight reported *p*-values. But even with access to the raw data, evaluating the evidence for the *substantive* inference would be impossible because the connection between data and claim is underspecified: It is unclear which aspects of the sparsely described method, coding procedure, or sampling strategy are necessary or sufficient to elicit informative data (although the need to avoid certain confounds is discussed), and exactly which pattern of data produced in this way would count as evidence for newborn imitation—and, importantly, which would not. In this, the article closely resembles most empirical publications that I have analysed in my own meta-scientific projects.

Preliminary operationalizations and fuzzy inferences are not a crime, but a normal starting point of scientific discovery. Yet in order to progress toward precise claims, the initial vagueness must be recognized and tackled in subsequent studies. Rozin (2001) described an idealized progression of a research line as follows:

> A first step is often verification that the phenomenon actually occurs. This may often be followed by an attempt to explore the generality of the phenomenon. A more disciplined description or exploration of the phenomenon often then ensues, with an attempt to discover laws or invariances. Such ventures are often not theory motivated, but rather are motivated by an attempt to be precise about the world, with the idea in mind that future theories will have to have something to explain. (p. 5)

If this were the normal course of events in psychology, most published articles should contain explicit efforts to refine existing concepts and measures rather than introducing new ill-defined claims. Meltzoff and Moore's (1977) study received immediate methodological and empirical pushback (Anisfeld, 1979; Jacobson & Kagan, 1979; Masters, 1979), but this did not lead the field towards a shared empirical specification of newborn imitation. The remarkable result is that over 40 years of research and disagreement (e.g., Anisfeld, 1996; Jones, 2009;

Meltzoff, 2017; Oostenbroek et al., 2016; Ray & Heyes, 2011) have failed to produce a consensus on whether the phenomenon actually exists (Davis et al., 2021).

3 | FIGHTS ABOUT FACTS

A recent failure to replicate newborn imitation (Oostenbroek et al., 2016) has been met with a slew of methodological and analytical criticism by the phenomenon's proponents (Meltzoff et al., 2018), some of which contradict those proponents' own previous arguments (Oostenbroek et al., 2019). As replication studies have gained popularity in recent years, this pattern of responses to failed replications has become a familiar sight. All across the discipline, proponents and sceptics of certain claims get stuck in seemingly endless back-and-forths of methodological criticisms and reanalyses without reaching the common ground (see, for example, the controversies on whether screen-time affects children's well-being, Kaye, Orben, Ellis, Hunter, & Houghton, 2020; Orben & Przybylski, 2019; Twenge, Haidt, Joiner, & Campbell, 2020; or on the hypothesis that hormonal changes during women's ovulatory cycles cause shifts in mate preference, Engber, 2018; Gangestad, Dinh, Grebe, Del Giudice, & Emery Thompson, 2019; Jünger, Kordsmeyer, Gerlach, & Penke, 2018). When different experts struggle so persistently to agree on what data would inform a research question, one may wonder if they are actually talking about the same thing—or, indeed, if there really is a question that can be answered with data. Dahl (2019) made a similar observation in a discussion of 'theoretical stalemates' in the field of moral development, noting that 'the accrual of data has not resolved major theoretical debates about how children develop moral concerns' and that 'it is unclear what evidence could resolve them' (p. 3). If data cannot speak to the scientific claims in the literature, what do such claims even *mean*?

4 | REPERCUSSIONS FOR SCIENTIFIC REFORM

Reforms proposed in response to the replication crisis have focused on improving the quality of empirical evidence by restricting 'researcher degrees of freedom' in the collection, analysis, and presentation of data (Simmons et al., 2011), for example, via preregistration (Nosek, Ebersole, DeHaven, & Mellor, 2018) or Registered Reports (Chambers & Tzavella, 2021), and by making it easier to verify published results, for example, via open data (Morey et al., 2016; Nosek & Bar-Anan, 2012), or improved computational reproducibility (Clyburne-Sherin, Fei, & Green, 2019; Hardwicke et al., 2018). But if content and meaning of scientific claims are so unclear that data cannot inform them, better evidence will fail to have an impact. Perhaps unsurprisingly, studies into the efficacy of preregistration have (unwittingly) shown that this new practice does not solve the problem of ill-defined claims and hypotheses. For example, Bakker et al. (2018) noted that even determining the exact *number* of hypotheses in a given preregistration had been so difficult that inter-rater agreement on this variable was as low as 14%. Similarly, Claesen, Gomes, Tuerlinckx, and Vanpaemel (2021) stated that '[a]ssessing the adherence of the published studies to the preregistration plans proved to be a far from trivial task' because 'neither the preregistration plans nor the published studies were written in sufficient detail for a fair comparison' (p. 4), and van den Akker (2021) remarked that '(r)esearchers are very bad at clearly laying out hypotheses in preregistrations (and in papers)' (p. 31).

These findings show that preregistration is no panacea for the epistemic problems of the discipline. Although formulating predictions and methods ahead of time can help identify underspecified parts of one's research plan, current implementations of the practice do not guarantee such insights.[2] In the worst case, preregistering ill-defined hypotheses may give researchers a false sense of precision. This way, the current focus on increased methodological rigour may lead us to overlook the need for more fundamental reforms. Instead of encouraging researchers to preregister precise but arbitrary specifications of ill-defined hypotheses, we should discuss the need for more

conceptual and theoretical work (Scheel, Tiokhin, Isager, & Lakens, 2021). Without this, investing in increasingly 'high-quality' data will not result in high-quality inferences.

5 | WHY THE FUZZINESS OF SCIENTIFIC CLAIMS CAN GO UNNOTICED

How is it possible that the above-mentioned problems encountered by meta-researchers went unnoticed by the authors, reviewers, and editors of the respective papers? A form of confirmation bias could be at play: Researchers may tend to merely ask themselves if a given operationalization or test result is *consistent with* a verbal claim, but not if it is *critical* for the claim. For meta-researchers attempting to analyse the evidence for a set of claims, the match between preregistration and a published article, or the reproducibility of a meta-analysis, the role is reversed: Because they are trying to verify something that has already been done, they need to identify not just *any* statement of a claim or result, but the *correct* one.

This may also explain the recent finding that asking different teams of researchers to test the same hypothesis—either by letting them analyse the same data (Botvinik-Nezer et al., 2020; Breznau et al., 2021; Silberzahn et al., 2018 or by letting each team devise its own operationalization (Landy et al., 2020)—produces approximately as many different conclusions as there are research teams involved. The four studies mentioned here guaranteed co-authorship to each participating researcher, thus eliminating an important incentive to 'exploit' researcher degrees of freedom. Although some participants may have been motivated to obtain results fitting their theoretical or ideological allegiances, a more mundane explanation is that the provided hypotheses were ill-defined and the researchers took more or less random walks through the multiverse of specifications left open by the study authors. This interpretation is supported by a recent reanalysis of Silberzahn et al. (2018) which attributes the lion's share of variation in outcomes between the 'many analysts' to them answering different versions of the underspecified research question (Auspurg & Brüderl, 2021).

6 | HOW CAN WE MOVE FORWARD?

To produce scientific claims that are less elusive and more meaningful, we need to recognize the broken parts of our inference chains and then try to repair them. The first two of the three broad recommendations below may help researchers to identify underspecified elements in their research, and the third is aimed at strengthening those weak elements. None of these suggestions is a silver bullet for what appears to be an entrenched problem of the discipline, but they may provide useful starting points.

1. **Formal modelling.** A key aspect of the problems described in this manuscript is that psychologists predominantly use verbal statements to express their theories, hypotheses, predictions, and inferences. Because natural language is imprecise, this practice keeps causing confusion. Expressing theoretical assumptions and hypotheses in formal mathematical, computational, or causal models can help reveal ambiguous definitions, hidden assumptions, and internal inconsistencies (Farrell & Lewandowsky, 2010; Frankenhuis & Tiokhin, 2018; Guest & Martin, 2021; Rohrer, 2018). An example can be found in Fenneman and Frankenhuis (2020), who used this approach to study whether the development of impulsive behaviour should be expected in harsh and unpredictable environments.
2. **Machine-readable hypothesis tests.** Whenever one wants to test a prediction using inferential statistics, the involved variables, sample, and evaluation criteria should be specified in unambiguous, standardized codes that allow a computer to evaluate the prediction once the data are in (Lakens & DeBruine, 2021). This procedure ensures a well-defined 'empirical reference' (de Groot, 1969) of the hypothesis. Widespread adoption of machine-readable hypothesis tests would also complement recent efforts of developmental researchers to build dynamic, community-augmented meta-analyses (Tsuji et al., 2017; Tsuji, Bergmann, & Cristia, 2014). An

accessible tutorial for making hypothesis tests machine-readable, including a software package and published examples, is provided by Lakens and DeBruine (2021). Helpful instructions for how to move from a hypothesis to an empirical test also can be found in de Groot (1969, esp. chapter 5).

3. **Nonconfirmatory research activities.** The practices listed above may reveal that one's approach was 'prematurely formal' (Rozin, 2001, p. 4). When researchers struggle to formalize their assumptions, or when the necessary methodological specifications seem exceedingly arbitrary, the way forward may be to take a step back and consider nonconfirmatory research activities to refine concepts, improve measurement, or identify auxiliary assumptions and boundary conditions (Scheel, Tiokhin, et al., 2021). These activities include descriptive research (qualitative and quantitative, naturalistic and lab-based), parameter-range exploration, and exploratory experimentation (see also Eronen & Bringmann, 2021; Yarkoni, 2020). Such approaches are by no means unfamiliar to developmental psychologists, but they deserve a (much) more prominent seat in the discipline's methodological repertoire. One exciting example is the use of head-mounted cameras to study the visual input that infants receive under naturalistic conditions (Aslin, 2009; Fausey, Jayaraman, & Smith, 2016).

7 | CONCLUSION

I have argued that many scientific claims in psychology are not sufficiently well-defined to be empirically testable. Although my argument implies that attempts to quantify the number of false claims in the literature may be misguided, this does not mean that there is no crisis. If anything, the situation might be worse than current replicability estimates imply—I suggest that most psychological research claims are *not even wrong*. This situation cannot be resolved with increased rigour in data collection and analysis. Instead, we need to go back to the drawing board and reconsider the conceptual basis of our research questions. In developmental research, first steps in this direction have already been taken: The *Many Babies Consortium* has established itself as a multi-lab collaboration that uses its resources to not simply scale up replication attempts of existing research, but to first establish consensus on conceptual definitions and experimental operationalizations (Visser et al., 2021). With some luck, the impetus of this important project and similar initiatives may lead us to a point where we can legitimately state that most psychological research findings are actually wrong—or, perhaps, slightly fewer than most (one can dream).

ENDNOTES

[1] Since 'finding' sounds deceptively objective and definitive, 'claim' may be the better option.

[2] One exception may be Registered Reports, in which the specificity and stringency of the research plan is evaluated by reviewers and editors. Initial evidence suggests that methodological quality and fit between method and research question are indeed greater in Registered Reports (Soderberg et al., 2021). However, even though they may reduce the problem, my own experience (e.g., Scheel, Schijen, & Lakens, 20212021) suggests that Registered Reports still contain too many ill-defined claims.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Anne M. Scheel https://orcid.org/0000-0002-6627-0746

REFERENCES

Anisfeld, M. (1979). Interpreting "imitative" responses in early infancy. *Science*, *205*(4402), 214–215. https://doi.org/10.1126/science.451593

Anisfeld, M. (1996). Only tongue protrusion modeling is matched by neonates. *Developmental Review*, *16*(2), 149–161. https://doi.org/10.1006/drev.1996.0006

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, *86*(6), 561–565. https://doi.org/10.1097/OPX.0b013e3181a76e96

Auspurg, K., & Brüderl, J. (2021). Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "many analysts one data set" project. *Socius*, *7*. https://doi.org/10.1177/23780231211024421

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D. T., & Wicherts, J. (2018). Ensuring the quality and specificity of preregistrations. https://doi.org/10.31234/osf.io/cdgyh

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Alvarez-Benjumea, A., … Nguyen, H. H. V. (2021). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. https://doi.org/10.31222/osf.io/cd5j9

Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex*, *49*, 606–610. https://doi.org/10.1016/j.cortex.2012.12.016

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of registered reports. *Nature Human Behaviour*, *1–14*. https://doi.org/10.1038/s41562-021-01193-7

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*, 211037. https://doi.org/10.1098/rsos.211037

Clyburne-Sherin, A., Fei, X., & Green, S. A. (2019). Computational reproducibility via Containers in Psychology. *Meta-Psychology*, *3*, 1–9. https://doi.org/10.15626/MP.2018.892

Dahl, A. (2019). Chapter one—the science of early moral development: On defining, constructing, and studying morality from birth. In J. B. Benson (Ed.), *Advances in child development and behavior* (pp. 1–35). Cambridge, MA: Academic Press. https://doi.org/10.1016/bs.acdb.2018.11.001

Davis, J., Redshaw, J., Suddendorf, T., Nielsen, M., Kennedy-Costantini, S., Oostenbroek, J., & Slaughter, V. (2021). Does neonatal imitation exist? Insights from a meta-analysis of 336 effect sizes. *Perspectives on Psychological Science*, *16*, 1373–1397. https://doi.org/10.1177/1745691620959834

de Groot, A. (1969). Methodology: Foundations of inference and research in the behavioral sciences. *Mouton*. Psychological Studies - Mouton (Vol. 6, pp. 1–400). The Hague, Netherlands: Mouton & CO.

Edelsbrunner, P., & Thurn, C. (2020). Improving the utility of non-significant results for educational research. https://doi.org/10.31234/osf.io/j93a2

Engber, D. (2018). The wax and wane of ovulating-woman science. *Slate Magazine* Retrieved from https://slate.com/technology/2018/10/ovulation-research-women-replication-crisis.html

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placì, S., Troisi, C. A., … Ostojić, L. (2020). Trialling meta-research in comparative cognition: Claims and statistical inference in animal physical cognition. *Animal Behavior and Cognition*, *7*(3), 419–444. https://doi.org/10.26451/abc.07.03.09.2020

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*(5), 329–335. https://doi.org/10.1177/0963721410386677

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107. https://doi.org/10.1016/j.cognition.2016.03.005

Fenneman, J., & Frankenhuis, W. E. (2020). Is impulsive behavior adaptive in harsh and unpredictable environments? A formal model. *Evolution and Human Behavior*, *41*(4), 261–273. https://doi.org/10.1016/j.evolhumbehav.2020.02.005

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., & Kruger, A. (2018). The epistemic importance of establishing the absence of an effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244. https://doi.org/10.1177/2515245918770407

Frankenhuis, W. E., & Tiokhin, L. (2018). Bridging evolutionary biology and Developmental Psychology: Toward an enduring theoretical infrastructure. *Child Development*, *89*(6), 2303–2306. https://doi.org/10.1111/cdev.13021

Gangestad, S. W., Dinh, T., Grebe, N. M., Del Giudice, M., & Emery Thompson, M. (2019). Psychological cycle shifts redux, once again: Response to Stern et al., Roney, Jones et al., and Higham. *Evolution and Human Behavior*, *40*(6), 537–542. https://doi.org/10.1016/j.evolhumbehav.2019.08.008

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10.1037/h0076157

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., … Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, *5*, 180448. https://doi.org/10.1098/rsos.180448

Ioannidis, J. P. A. (2005). Why Most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jacobson, S. W., & Kagan, J. (1979). Interpreting "imitative" responses in early infancy. *Science*, *205*(4402), 215–217. https://doi.org/10.1126/science.451594

Jones, S. S. (2009). The development of imitation in infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2325–2335. https://doi.org/10.1098/rstb.2009.0045

Jünger, J., Kordsmeyer, T. L., Gerlach, T. M., & Penke, L. (2018). Fertile women evaluate male bodies as more attractive, regardless of masculinity. *Evolution and Human Behavior*, *39*(4), 412–423. https://doi.org/10.1016/j.evolhumbehav.2018.03.007

Kaye, L. K., Orben, A., Ellis, D. A., Hunter, S. C., & Houghton, S. (2020). The conceptual and methodological mayhem of "screen time". *International Journal of Environmental Research and Public Health*, *17*(10), 3661. https://doi.org/10.3390/ijerph17103661

Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable: *Advances in methods and practices in psychological*. *Science*, *4*. https://doi.org/10.1177/2515245920970949

Lakens, D., Page-Gould, E., van Assen, M. A. L. M., Spellman, B., Schönbrodt, F., Hasselman, F., Corker, K. S., Grange, J. A., Sharples, A., Cavender, C., Augusteijn, H. E. M., Augusteijn, H., Gerger, H., Locher, C., Miller, I. D., & Anvari, F. (2017). Examining the reproducibility of meta-analyses in psychology: A preliminary report. *MetaArXiv*. https://doi.org/10.31222/osf.io/xfbjf

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., … The Crowdsourcing Hypothesis Tests Collaboration. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*, 451–479. https://doi.org/10.1037/bul0000220

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS One*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

Meltzoff, A. N. (2017). Elements of a comprehensive theory of infant imitation. *Behavioral and Brain Sciences*, *40*, e396. https://doi.org/10.1017/S0140525X1600193X

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human *neonates*. *Science*, *198*(4312), 75–78. https://doi.org/10.1126/science.198.4312.75

Meltzoff, A. N., Murray, L., Simpson, E., Heimann, M., Nagy, E., Nadel, J., … Ferrari, P. F. (2018). Reexamination of Oostenbroek et al. (2016): Evidence for neonatal imitation of tongue protrusion. *Developmental Science*, *21*(4), e12609. https://doi.org/10.1111/desc.12609

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … Zwaan, R. A. (2016). The peer Reviewers' openness initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*, 150547. https://doi.org/10.1098/rsos.150547

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*(3), 217–243. https://doi.org/10.1080/1047840X.2012.692215

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., … Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1) annurev-psych-020821-114157. https://doi.org/10.1146/annurev-psych-020821-114157

Oostenbroek, J., Redshaw, J., Davis, J., Kennedy-Costantini, S., Nielsen, M., Slaughter, V., & Suddendorf, T. (2019). Re-evaluating the neonatal imitation hypothesis. *Developmental Science*, *22*(2), e12720. https://doi.org/10.1111/desc.12720

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., … Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology*, *26*(10), 1334–1338. https://doi.org/10.1016/j.cub.2016.03.047

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Ray, E., & Heyes, C. (2011). Imitation in infancy: The wealth of the stimulus. *Developmental Science*, *14*(1), 92–105. https://doi.org/10.1111/j.1467-7687.2010.00961.x

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2). https://doi.org/10.1177/25152459211007467

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

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. https://doi.org/10.1177/0956797611417632

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., … Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, *5*(8), 990–997. https://doi.org/10.1038/s41562-021-01142-4

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*(1), 108. https://doi.org/10.2307/2684823

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665. https://doi.org/10.1177/1745691614552498

Tsuji, S., Bergmann, C., Lewis, M., Braginsky, M., Piccinini, P., Frank, M. C., & Cristia, A. (2017). MetaLab: A repository for meta-analyses on language development, and more. *Interspeech*, *2017*, 2 Retrieved from https://www.isca-speech.org/archive/Interspeech_2017/pdfs/2053.PDF

Twenge, J. M., Haidt, J., Joiner, T. E., & Campbell, W. K. (2020). Underestimating digital media harm. *Nature Human Behaviour*, *4*(4), 346–348. https://doi.org/10.1038/s41562-020-0839-4

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2021). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. https://doi.org/10.31234/osf.io/af9by

van den Akker, O. (2021). *Selective hypothesis reporting in psychology* [conference presentation]. 9th BITSS Annual Meeting [virtual conference]. Retrieved from https://osf.io/c6rnb/

Visser, I., Bergmann, C., Byers-Heinlein, K., Ben, R. D., Duch, W., Forbes, S. H., & Zettersten, M. (2021). Improving the generalizability of infant psychological research: The Many Babies model. *PsyArXiv*. https://doi.org/10.31234/osf.io/8vwbf

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685

How to cite this article: Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), e2295. https://doi.org/10.1002/icd.2295