Commentary

Dan Dediu*, Maria Koptjevskaja-Tamm and Kaius Sinnemäki

Replication, robustness and the angst of false positives: a timely target article and its multifaceted comments

https://doi.org/10.1515/lingty-2025-0065
Published online July 30, 2025

As scientists, we[1] love what we are doing and we hope that our findings will outlive us, and we dread being shown wrong (probably for many of us, nightmares of discovering we're naked in public are far outranked by being shown wrong in public). This is a foundational paradox, as any first year undergrad will breathlessly and smugly tell you that all theories are false and that what sets science apart from everything else is its falsifiability (Godfrey-Smith 2021; Newton-Smith 2001; Psillos and Curd 2008). Which means, by extension, that everything we do is ultimately wrong.

But surely there are degrees of wrongness: how many of us would not wish to be wrong the way Newton was, given that most of what we do, build, and even launch in space is still based on his theory, spectacularly falsified by a certain Albert more than a century ago. Or the way Darwin got it wrong with inheritance. Some kinds of wrongness are to be desired at all costs, as they make us advance, question fundamental assumptions, and find new, previously invisible paths. Other kinds of wrongness, such as fundamental overlooked problems in data analysis, are deeply dreaded. Another foundational paradox is the following: in general we dread being wrong but there can be no advancement without being wrong: discovery is, despite what various voices keep claiming, messy, costly, wasteful and resists "optimisation",

1 Here, we (the authors) use this "we" to refer to all of us, the scientists.

***Corresponding author: Dan Dediu [dan ˈdedju]**, Department of Catalan Philology and General Linguistics, University of Barcelona, Barcelona, Spain; University of Barcelona Institute for Complex Systems (UBICS), Barcelona, Spain; and Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain, E-mail: dan.dediu@icrea.cat
Maria Koptjevskaja-Tamm [ma ˈ ria kɔpˈtʃɛfskaja ˈtam], Department of Linguistics, Stockholm University, Stockholm, Sweden, E-mail: tamm@ling.su.se. https://orcid.org/0000-0002-9592-5780
Kaius Sinnemäki [ˈkaius ˈsin:emæki], General Linguistics, University of Helsinki, Helsinki, Finland, E-mail: kaius.sinnemaki@helsinki.fi. https://orcid.org/0000-0002-6972-5216

"rationing" and "ideology". Discovery is intrinsically built on getting it wrong again, and again, and again.

Simplifying in the extreme, we can get it wrong in two ways (Peterson 2009): claim that something is when it ain't (a so-called FALSE POSITIVE or type I error) or the other way around, that there's nothing interesting to see when, in fact, there is (FALSE NEGATIVE or type II error). For lots of reasons, it is the false positives that seem to keep most people awake at night, and the fear of those is inculcated in every student that ever sat in a statistics class where hypothesis testing and *p*-values were repeatedly drilled into them.[2] And this is why a study needs to be repeated as well, as there are lots of ways null effects can still masquerade as "significant" findings[3] – and this is precisely where our target article, *Replication and methodological robustness in quantitative typology*,[4] comes in.

We will not summarise it, nor the many very relevant comments it has generated, as we would rather leave the pleasure of drawing their own conclusions to our diverse readership, but instead, briefly discuss why we decided to host this debate in *Linguistic Typology*. First, our endeavour is eminently EMPIRICAL and interested in the real, messy, complex, always surprising world of language and languages, which means that most of our claims are potentially false, and not in the ways of Newton or Darwin, but in the much more boring and common ways of false claims that have generated the various "replicability crises" (Ioannidis 2005; Vasishth et al. 2018). We are unwilling to add to those crises a "typological" one, but, given the traditional small samples, lack of methodological agreement and tendency to draw grand conclusions, we suspect that there might, indeed, be one, despite the increased attention given to these issues over the years, in particular, in *Linguistic Typology* (e.g., Jaeger et al. 2011; Janssen et al. 2007; Editorial Board 2016). Second, our field is fast becoming heavily QUANTITATIVE, methodologically very sophisticated, and with access to large databases, which makes it ready to embrace more formal ways of buttressing its claims. Third, whether we like it or not, our field is IMPORTANT, in that the patterns of linguistic diversity, their causes and effects affect many human enterprises, generate genuine interest and even have actual consequences in the real world of money, power and justice, forcing us to be extra careful with our claims. Finally, the submitted paper was, frankly, very good and inciting, irrespective of whether one agrees or not with its claims, and obviously, in need of further replication.

---

2 Bayesian statisticians tend to be less dogmatic (Gelman et al. 2020; Kruschke 2014; McElreath 2020), but still…

3 Small samples are particularly prone, but biased designs, problematic samples, unjustified assumptions and many other factors conspire to fool even the best of us (Button et al. 2013; Cohen 1988; Gelman and Carlin 2014; Ioannidis 2005; Kraft 2008; Vasishth et al. 2018).

4 Becker, Laura and Guzmán Naranjo, Matías. 2025. Replication and methodological robustness in quantitative typology. *Linguistic Typology*. https://doi.org/10.1515/lingty-2023-0076

Before ending, we want to briefly return to the false claims nobody seems to care about: the NEGATIVE ones. They matter, too, and guide research along with positive ones, although they may not be as visible in practice. Overlooking an intriguing fact because it did not reach "significance", passing by a tiny, overgrown path in the forest, or ignoring a furtive look in the subway, could be missed opportunities to cure cancer, miss a breathtaking view, or fail to find the love of your life.[5] More to the point: because of the perverse way the world works, we cannot simultaneously decrease the chances of a false negative AND of a false positive (Kim 2015; Nakagawa et al. 2024): the more afraid we are of making false discoveries, the more we will miss genuine ones (Peterson 2009). It's a choice we must make as a scientific field and as a society. What hurts more (and when): misleading and directing scarce resources into avenues that don't exist, or trudging on well-worn motorways and missing the little diverging paths? Given the inevitability of having to make this choice over and over again under complex, explicit and implicit, conscious and unconscious, economic, ethical, scientific, ideological, etc., etc., etc. constraints, we are looking forward to how our scientific field develops amidst this tug of war.

References

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5). 365–376.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral science*, 2nd edn. New York: Routledge.

Editorial Board. 2016. Re-doing typology. *Linguistic Typology* 10(1). 67–128.

Gelman, Andrew & John Carlin. 2014. Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9(6). 641–651.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao & Lauren Kennedy. 2020. Bayesian workflow. *arXiv*. https://doi.org/10.48550/arXiv.2011.01808.

Godfrey-Smith, Peter. 2021. *Theory and reality: An introduction to the philosophy of science*, 2nd edn. Chicago: University of Chicago Press.

Ioannidis, John P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8). e124.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15. 281–319.

Janssen, Dirk P., Balthasar Bickel & Fernando Zúñiga. 2007. Randomization tests in language typology. *Linguistic Typology* 10(3). 419–440.

Kim, Hae-Young. 2015. Statistical notes for clinical researchers: Type I and type II errors in statistical decision. *Restorative Dentistry & Endodontics* 40(3). 249–252.

5 It's not just the three of us being from the "analogic" generations, but even the Spanish singer *El Kanka* (https://es.wikipedia.org/wiki/El_Kanka) thinks the same ("Cuando el destino llamó a tu puerta tenías puesto los auriculares. Pudo pasar, pero no lo escuchaste" in "Pudo Pasar").

Kraft, Peter. 2008. Curses – Winner's and otherwise – in genetic epidemiology. *Epidemiology (Cambridge, Mass.)* 19(5). 649–651. Discussion 657–658.

Kruschke, John. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Cambridge, Massachusetts: Academic Press.

McElreath, Richard. 2020. *Statistical rethinking: A Bayesian course with examples in R and stan*. New York: CRC Press.

Nakagawa, Shinichi, Malgorzata Lagisz, Yefeng Yang & Szymon M. Drobniak. 2024. Finding the right power balance: Better study design and collaboration can reduce dependence on statistical power. *PLOS Biology* 22(1). e3002423.

Newton-Smith, W. H. 2001. A companion to the philosophy of science. US: Wiley.

Peterson, Martin. 2009. *An introduction to decision theory (Cambridge Introductions to Philosophy)*. Cambridge: Cambridge University Press.

Psillos, Stathis & Martin Curd. 2008. *The Routledge companion to philosophy of science*. New York: Routledge.

Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger & Andrew Gelman. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.