Caterina Mauri and Andrea Sansò*

Replicability all the way up: commentary on "Replication and methodological robustness in quantitative typology" by Becker and Guzmán Naranjo

https://doi.org/10.1515/lingty-2025-0033
Received April 7, 2025; accepted May 21, 2025; published online July 17, 2025

Becker and Guzmán Naranjo (2025) (henceforth B&GN) present a compelling case study demonstrating that applying different statistical analysis techniques to the same dataset can yield divergent conclusions, particularly when these techniques account for potential genealogical and areal biases overlooked in the original study. They argue that replicability constitutes a crucial added value for any large-scale study and should be actively pursued as a key objective. Ensuring replicability, they suggest, is one of the most effective strategies for guaranteeing that linguistics can address both longstanding and emerging questions with methodological robustness.

B&GN specifically operationalize replicability as the possibility of applying different statistical models to data analysis. However, their broader definition of replicability implicitly extends to all stages of analysis, from data collection to classification and labeling. Notably, in the appendix of their study, B&GN devote a specific section to guidelines for the transparent documentation of the language sample. This section covers both the sampling procedure and the explanation of the annotation choices. Yet, they themselves acknowledge that, apart from Nichols et al. (2006), there are no studies on replicability that systematically examine how different classificatory choices or the use of alternative parameters impact data analysis and annotation.

This is the point on which we would like to focus. It highlights the importance of shifting attention further upstream to a prerequisite that precedes replicability and transparency in statistical model selection. While this aspect has already been implicitly addressed in many studies, the scientific community has yet to engage in a

***Corresponding author: Andrea Sansò [an'drea san'sɔ]**, Università di Salerno, Fisciano, Italy,
E-mail: asanso@unisa.it
Caterina Mauri [kate'rina 'mawri], Università di Bologna, Bologna, Italy,
E-mail: caterina.mauri@unibo.it

collective discussion on how best to implement it. Thus, in this commentary, we aim to address several issues that B&GN leave in the background. We believe these issues are equally important for advancing the broader goals of replicability, robustness and open science.

A second point, closely related to the first, that we wish to briefly discuss in this commentary concerns the fact that in many cases of quantitative studies, what matters is not so much (or at least not only) replicability in the statistical sense described by B&GN, but rather the ability to critically examine (and, if necessary, revise) the theoretical foundations of the quantitative investigation.

In the following sections, we will discuss these two points in detail (see Sections 1 and 2, respectively). In Section 3, we will present a reflection aligned with this perspective, focusing on a research project we are currently engaged in. This project investigates the potential impact of literacy on grammatical structures and the methodological challenges associated with such an inquiry.

1 Why it is important to ensure replicability in data collection and classification

The practice of making typological datasets publicly accessible has fortunately become increasingly widespread. Researchers typically verify datasets carefully before release. As B&GN note in their appendix, such datasets should ensure transparency regarding the methodological choices that led to a particular classification of the data.

In this regard, we highlight the comment field model in Grambank (Skirgård et al. 2023) as an example of best practice. Recording the rationale for a given classification, this field is particularly valuable because it allows researchers to revise annotations if new data emerge or if their classification criteria differ from those of the Grambank compilers. In contrast, resources like WALS (Dryer and Haspelmath 2013), though highly valuable, remain somewhat opaque, often leaving users uncertain about the reasoning behind a given classification.

We propose that B&GN's guidelines also include this recommendation: once theoretical and methodological choices are clearly stated, datasets should offer a computationally simple way to reverse or modify them – reversing binary or ternary classifications (like "Y/N/?"), and modifying multi-valued categories. This is not simply a matter of coding hygiene: it is a precondition for enabling rigorous and cumulative science. While this may seem a simple, even trivial, task, its importance should not be underestimated. Otherwise, biases or misrepresentations introduced at the outset risk being carried through the statistical analysis. These biases may often seem like mere background noise, but can we really be sure that is always the case?

2 Replicability, common sense, and the need for a social layer

An important observation in this regard is that while replicability in statistical modeling is crucial for examining the same data from different perspectives, there are cases – such as those noted by B&GN (e.g., their discussion of Chen 2013) – where replicating the statistical analysis is unnecessary to recognize fundamental theoretical weaknesses. In Chen's study, the hypothesis of a statistical link between future tense marking and a society's propensity for saving is poorly formulated for several reasons; the fact that the absence of a dedicated future marker does not prevent speakers from referring to the future, or the reliance on a naïve interpretation of linguistic relativity.

The same applies to other, more sophisticated quantitative studies. For instance, Blasi et al. (2019) propose that the shift from a hunter-gatherer to an agricultural society led to an increase in labiodental sounds. While replicating this study with alternative statistical methods might yield valuable insights – open commentaries to Blasi et al. (2019) suggest that accounting for factors like distance from Africa (Tarasov and Uyeda 2020) or consonant inventory size (Berthommier and Boë 2019) produces different results[1] – the study's core assumptions remain difficult to substantiate. As Huijbregts (2019, 2020) points out, for instance, changes in bite configuration did not possibly alter the universal phonetic capabilities of humans, which remained unchanged: labiodental sounds do not constitute an evolutionary innovation, as the phonetic features required to produce them were already available in pre-Neolithic human speech systems.[2] Other significant methodological limitations can be formulated as follows: First, can we reliably ascribe phonetic substance to the reconstructed phonemic inventories of proto-languages? Likely not. These reconstructions are better understood as symbolic representations of systematic correspondences between attested languages – a kind of "warranty" of genetic relatedness rather than actual phonological systems. Second, even if we assume that a reconstructed phonological system reflects a real linguistic entity, the pervasive nature of phonetic variation in living languages is an aspect of the Implicit Uniformitarian Assumption (as

1 For the sake of completeness, it should be noted that Blasi et al. (2020) respond to analyses based on alternative factors – such as distance from Africa – by showing that the claims made by Tarasov and Uyeda (2020) are affected by issues of interpretation, data selection, and lack of genealogical controls, and that their analyses do not adequately account for confounding variables.

2 Moran and Bickel (2020) reply to Huijbregts (2020) criticizing him for failing to engage with empirical data and for relying on outdated phonological theories, such as the *Sound Patterns of English* system of distinctive features (cf. Chomsky and Halle 1968), which they argue inadequately accounts for many languages and phonological phenomena.

formulated and criticized by Moran et al. 2021) that cannot be easily dispensed with. This raises an important question: what ensures that phonemes in the hypothetical proto-language speech community were not realized with considerable phonetic diversity?

It is important to clarify that the aim here is not to diminish the value of Blasi et al.'s quantitative work but rather to emphasize the necessity of two further dimensions of scientific transparency: one is the opportunity for open commentaries on quantitative studies, a practice increasingly adopted by academic journals, but perhaps not so common in linguistics (and linguistic typology) journals. This approach fosters a critical discourse that extends beyond mere replicability, ensuring that both the potential insights of a given classification and its possible limitations can be openly examined and debated. The other one is the opportunity for datasets in typological research to incorporate a social layer, understood, following Cysouw (2007: 63), as an annotation layer intended to foster collaboration and discussion. This layer should minimally include information relevant to a specific research question and to the process of annotation and classification, allowing others interested in the same (or a closely related) topic to easily access the data and draw their own conclusions. Only the incorporation of such a social layer into scholarly endeavors will enhance the robustness of scientific inquiry and contribute to a more dynamic and self-correcting research environment.

3 A case study: investigating the impact of literacy on grammatical structures

In this final section, we draw on an ongoing research project to highlight the importance of implementing a social layer and ensuring methodological transparency, and thus replicability, at the upstream stages of research. These are the stages at which theoretical constructs are operationalized, and data is selected, annotated, and interpreted. Our study on the possible effects of literacy on grammatical structures (Mauri et al. 2025) offers a concrete example of why replicability must extend beyond statistical modeling to include the earlier, often less visible, phases of the research process, together with the challenges this entails.

A major difficulty we have encountered in the very first stages of our research concerns the availability and consistency of literacy data in cross-linguistic repositories. Not only is the information often incomplete or out of date, but it is also based on unclear definitions of what counts as "literacy". Is literacy to be measured at the national level or at the level of specific language communities? Should it include second-language literacy (e.g., literacy only in a second, often

official language), and in this case, how could we disentangle its effects? What thresholds of reading and writing proficiency should be considered relevant for evaluating potential structural effects?

These questions are not purely technical, but they shape the very structure of the dataset and influence the formulation of the research questions. They also affect how we interpret co-occurrences between grammatical features and sociolinguistic variables. For instance, observed patterns in clause combining may correlate with the presence of literacy, but could also result from other sociopolitical processes such as standardization or education policy. This raises the question: are we studying the effects of literacy itself, or of the broader historical and ideological apparatus with which it tends to co-occur?

Our experience suggests that replicability in broad studies like ours, i.e., on phenomena that are somehow controversial, both at the definitory and operationalizing levels, cannot be achieved without full transparency regarding how the foundational theoretical decisions are taken and, crucially, without enabling others to RECONFIGURE the underlying assumptions and categorizations. Inspired by B&GN's emphasis on methodological robustness, we advocate for the implementation of a social layer within the dataset itself, allowing for annotations to be revisited, challenged, or recalibrated in light of new data or alternative criteria.

Finally, given the complexity of sociolinguistic variables like literacy, which require a deep and expert knowledge of linguistic communities, we see considerable potential in crowd-sourcing initiatives and community-based data collection. These approaches could both enrich the empirical base and embed the research process within a more dialogic and reflexive framework. But again, such strategies are only useful if they are underpinned by transparent documentation and by mechanisms that facilitate the collaborative critique and refinement of the data.

In sum, before applying any statistical model – whether simple or advanced – it is essential to ensure that the data and the theoretical scaffolding supporting it are themselves open to replication and revision. As our case study illustrates, methodological robustness must begin well before model fitting, and it depends as much on epistemological reflexivity as on technical replicability.

References

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

Berthommier, Frédéric & Louis-Jean Boë. 2019. Labiodental fricatives /f v/ are less present in languages having a small number of consonants. Electronic response to "Blasi et al. 2019, Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363. eaav3218.

Blasi, Damián E., Steven Moran, Scott R. Moisik, Widmer Paul, Dan Dediu & Balthasar Bickel. 2019. Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363(6432). https://doi.org/10.1126/science.aav3218.

Blasi, Damián E., Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel. 2020. Languages, evolution and statistics: Human sound systems were shaped by changes in bite configuration. Response to Tarasov & Uyeda (2020). *BioRxiv preprint*. https://doi.org/10.1101/2020.02.27.965400.

Chen, Keith. 2013. The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review* 103(2). 690–731.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Cysouw, Michael. 2007. A social layer for typological databases. In Andrea Sansò (ed.), *Language resources and linguistic theory*, 59–66. Milan: Franco Angeli.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. WALS Online (v2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591.

Huijbregts, Riny. 2019. Emergent labiodentals are no factor in language evolution. Electronic response to "Blasi et al. 2019, Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363. eaav3218.

Huijbregts, M. A. C. (Riny) 2020. Biting into evolution of language. *Journal of Language Evolution*. 175–183. https://doi.org/10.1093/jole/lzaa003.

Mauri, Caterina, Andrea Sansò, Silvia Ballarè & Ludovica Pannitto. 2025. How literacy shapes grammar: Quantitative insights and methodological challenges of a typological approach. Paper presented at the conference *Linguistic data and language comparison in light of the 'quantitative turn' and 'big data'*. Bern 7–9 May 2025.

Moran, Steven & Balthasar Bickel. 2020. Rejoinder to Huijbregts's: Biting into evolution of language. *Journal of Language Evolution*. 1–4. https://doi.org/10.1093/jole/lzaa005.

Moran, Steven, Nicholas A. Lester & Eitan Grossman. 2021. Inferring recent evolutionary changes in speech sounds. *Philosophical Transactions of the Royal Society B* 376. 20200198.

Nichols, Johanna, Jonathan Barnes & David Peterson. 2006. The robust bell curve of morphological complexity. *Linguistic Typology* 10(1). 96–106.

Skirgård, Hedvig, Hannah J. Haynie, Harald Hammarström, Damián Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage. 2023. Grambank v1.0 (v1.0) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.7740140.

Tarasov, Sergei & Josef Uyeda. 2020. 2 reasons to refute Hockett's hypothesis that bite configuration affects human sound evolution. Electronic response to "Blasi et al. 2019, Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363. eaav3218.