Commentary

Adam J. R. Tallman*

Weak theories and robustness: Commentary on "Replication and methodological robustness in quantitative typology" by Becker and Guzmán Naranjo

https://doi.org/10.1515/lingty-2025-0032
Received April 4, 2025; accepted May 30, 2025; published online July 30, 2025

Becker & Guzman-Naranjo[1] (B&GN) provide statistical (re)analyses of four previously published typological studies. Their (re)analyses are unified around their use of phylogenetic regression with a Gaussian process model for areal effects. However, the original studies they refer to and reassess use a wide variety of methodologies from visual assessment combined with qualitative historical knowledge (Seržant 2021), the relative ranking of adjusted frequency scores (Dryer 2018), to sophisticated phylogenetic models (Shcherbakova et al. 2023). Their article takes up three issues. First, statistical models can pick out more "complex patterns" than less statistically sophisticated methods of assessment. Secondly, statistical (re)analyses that use different modeling assumptions are important for assessing the robustness of the original claims of a study.[2] Thirdly, in at least one case (with Shcherbakova et al.), their discussion considers issues of statistical (mis)specification.

B&GN summarize a number of studies in typology where statistical (re)analyses (dis)confirmed or did not replicate the findings of the earlier study. In my view, their framing (at least implicitly) contextualizes their study in relation to the broader replication crisis. The replication crisis has resulted in researchers

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

[2] Perhaps more controversially they view their study as a species of "replication", despite the fact that they do not gather new data (i.e. they follow Schmidt 2009 and reject Machery 2020 in this sense). From what I can discern B&GN do not provide an argument in favor of their conflation of replication with robustness. Regardless of whether one follows the authors here, tests of robustness are widely regarded as important for assessing the validity of scientific claims (Wimsatt 2007).

***Corresponding author: Adam J. R. Tallman [ˈærm dʒeɪmz ras ˈtalmn]**, Friedrich-Schiller-Universitat Jena, Saxony, Germany, E-mail: adam.james.ross.tallman@uni-jena.de

providing methodological and statistical fixes to science more broadly. However, a number of authors have argued that the problems of replication and robustness relate to a deeper crisis of theoretical underdevelopment (Fried 2020; Muthukrishna and Henrich 2019; Szollosi et al. 2020; van Rooij and Baggio 2021, inter alia). I argue that this point is relevant for the analyses discussed by B&GN. Of specific relevance, Fried (2020) makes a distinction between "strong" and "weak" theories. Whereas the former are explicit about underlying assumptions, the latter are imprecise narratives.

Weak theories are more susceptible to problems of replication/robustness because they are plagued by "hidden assumptions and other unknowns". A statistical model imposes assumptions on the data that are supposed to be consistent with the theory that it models, but weak theories can be formalized into statistical models in a number of different ways. A specific symptom of weak theories might be that they do not predict precise effect sizes. Shcherbakova et al. (2023) seem to be a case in point since they only mine for trade-offs with no specific theory about how strong these should be.

In the case of weak theories, a lack of robustness might not be a question of the veracity of a given theory, but rather we might question whether we are sure what the theory articulated in the abstract actually means to begin with (see Smaldino 2017). Furthermore, if a statistical model meant to test the robustness of a prior claim introduces assumptions that are not in the original study, one could also claim that one is testing a different theory. The fact that the adoption of these assumptions is not ruled out in the theory discussed in the original study could be a sign of theoretical underdevelopment.

Whereas B&GN are concerned with the extent to which the results of their statistical (re)analyses align with the qualitative assessments or quantitative results of the research they reanalyze, I would turn the focus to the relationship between theoretical claims in the original papers and the statistical models employed by B&GN. A lack of robustness in statistical results might ultimately relate to the fact that the theories under consideration are "weak" in the sense described above.

The article by Seržant (2021) is based on visual exploration of the decay factor. The main causal factor is the particular contact configuration that is thought to hold between speech communities. The contact configuration refers to "different local factors such as the time depth of the contacts, the particular historical events, the physical shape of the territory (cf. Nichols 1992), and other factors" (p. 67). As the strength of contact between two languages increases, so should their decay factors converge towards a similar result. The visual data of the decay factor show an East-West cline (more decay as we go west), but Seržant states that the data show "skewing" in a way that requires additional reference to the specific contact configuration between Slavic language groups and Turkic and Uralic tribes. The decay rate is lower than it would be otherwise. In so doing, Seržant slips in a hidden assumption concerning what a normal non-skewed distribution would look like without making it quantitatively explicit. How can we test Seržant's claim without a theory about what the decay factor would look like without the contact configuration he discusses?

Although not explicitly addressed in this fashion, we could assume B&GN's Gaussian process model for areal effects could function as the implicit theory underlying Seržant's skewing observation. But the authors actually face an unacknowledged dilemma in assessing Seržant's hypothesis. The main explanatory factor in Seržant (2021) is a special contact situation between Slavic/Turkic/Uralic languages, but this is not coded in the data available to B&GN. In order to test whether the contact configuration is necessary, they would, in principle, have to build a model with and without the special contact configuration as a predictor.

They ignore the problem and conclude "[s]ince this area [the contact area between Slavic/Turkic/Uralic] is very large and reflects the absence of innovation, it can be taken as a default situation as opposed to the two hotbeds identified." But this assertion is not directly related to Seržant's theory since the contact configuration is excluded from consideration by not being coded in any of their models. B&GN assume that spatial relations by themselves can serve as proxies for all forms of contact (indeed, they do not have the data coded in such a way to do otherwise), an assumption which dismisses Seržant's hypothesis at the onset rather than testing it. The main challenge here is how to actually measure and code the contact configuration so that we can see whether we need more than simply an East-West cline (and genealogy) to test Seržant's theory. Another issue is, of course, how precisely Seržant's latent theory about the East-West cline, against which he notices a skewed distribution, should be constructed in the first place.

In contrast to Seržant, the article by Shcherbakova et al. (2023) uses computational phylogenetic modeling. Such coevolutionary models are used to assess correlations between the presence or absence of certain grammatical markers across verbal and nominal domains over time in relation to a phylogenetic model. Their article is concerned with trade-offs between features, but the relationship between the theoretical claims and the results of their models are not fully articulated and are, in some cases, difficult to discern. First, the authors state that the motivation underlying a trade-off between two grammatical domains relates to redundancy and efficiency (e.g. "some grammatical categories are marked on nouns as well verbs", Shcherbakova et al. 2023: 156). They mention two candidate trade-offs; (i) case marking on nouns and argument marking on the verb; (ii) case marking on nouns and tense on verbs. In both cases there are some conditions such that overlapping grammatical functions are encoded in both the nominal and verbal domain. However, in the last paragraph before data analysis they state they are searching for any correlations between any of the features in their data. One might argue that this fits into a broader theory concerning equi-complexity. However, in this case it is unclear why trade-offs between all grammatical features are not assessed.

Shcherbakova et al. (2023) find evidence for a negative trade-off between verbal and nominal domains in Indo-European (IE), but evidence for the co-development (anti-trade-off) of features in Sino-Tibetan, and no general trend in Austronesian. In IE, the result is contingent on the tree adopted. No obvious explanation for the lineage-specific effects they find is apparent and they find no evidence for coevolution between nominal and verbal grammatical coding at a global level. One latent assumption that is made by the model is that there are no interactions or relationships between any of the grammatical features within nominal and verbal domains (e.g. it is assumed there is no diachronic relationship between grammatical expression of tense and aspect), an assumption which seems rather implausible given what we know about grammaticalization trajectories in these domains (e.g. Bybee et al. 1994). They then conduct pairwise comparisons between nominal and verbal grammatical categories.

In so doing they make another assumption that each pairwise comparison reflects a direct relationship between the variables uncontaminated by other ("lurking") variables which could act as a common cause: it is assumed that none of the variables are confounds in the assessment of the relationship between any two correlated variables. They find positive and negative correlations between certain features, but theoretical explanations are not available. As such the analysis could (or should) be regarded as mostly exploratory.

B&GN perform a statistical reanalysis of Shcherbakova et al. (2023). Their results call into question the lineage-specific claim of Shcherbakova et al. (2023) because they find that the effect of genealogy generally disappears when spatial control is added in the model.[3]

In assessing the relationship between specific variables, they adopt a very different modeling strategy and include all the features in the model at once, justifying their approach as follows: "[b]y including all feature groups, the model has

[3] However, the comparison may be obscured by the fact that Shcherbakova et al. (2023) and B&GN model phylogenetic relatedness in different ways: the former uses a coevolutionary model (via BayesTraits), while the latter employs phylogenetic regression. Although both approaches can be seen as implementing a form of genealogical control in B&GN's sense, they rest on different assumptions about how correlations between traits relate to tree topology. In coevolutionary models, trait changes along each branch are sampled from a bivariate distribution, preserving the structure of the tree as a process through which traits evolve. In contrast, B&GN's phylogenetic regression simplifies the tree into a matrix of pairwise relationships (i.e., a covariance structure) among languages, based on their degree of relatedness. As a nonexpert in these models, my intuition is that for the two approaches to yield comparable results, several conditions would need to hold – one of which is that the tree would have to be "balanced", meaning the number of languages is evenly distributed across branches. In any case, we are given no reason to believe that the two studies are controlling for genealogy in equivalent – let alone theory-neutral – ways.

more information about how the different feature groups are associated with each other, and it can produce more reliable estimates".

However, this claim slips in the assumption (without argument) that none of the variables included in their model function as colliders between other variables. A collider variable C for variables A and B is one where A and B independently cause C. For instance, the presence of grammatical tense (probabilistically) causes grammatical aspect marking and grammatical case (probabilistically) causes grammatical aspect marking (or the absence of it) cross-linguistically and over time. (tense → aspect ← case). However, in assessing whether there is a correlation between tense and case, adding aspect as a predictor, might result in a spurious correlation between these variables (see McElreath 2020: Chapter 6; Pearl 2009). B&GN's no-collider assumption is not any better than Shcherbakova et al.'s (2023) assumption that there are no mediating variables. The point here is that statistical models can only be judged as better or worse in relation to a theoretical structure which maps out causal relations between variables. The problem is that the statistical assessments are being done without explicit reference to causal theories about the diachronic relationship between grammatical coding.

B&GN make an important contribution to linguistic typology by emphasizing the need for statistical tests of robustness. However, it is possible that their paper obscures a serious problem in typology. Typological theories may fail robustness tests not just because they are false or because we should be less certain about their veracity, but rather because they are weak theories, vulnerable to hidden assumptions. As such B&GN might undersell the importance of statistical model comparison in general, because it can provide us with important tools for highlighting the lack of clarity in our typological theories. We might agree on some general abstract idea but be completely misaligned concerning the details necessary to render our claims testable in a consistent manner.[4] Likewise, we might misunderstand the basis for our disagreements if we adopt different (hidden) assumptions. I would conclude by emphasizing the importance of developing explicit theories alongside methods for assessing the robustness of statistical effects (e.g. van Rooij and Baggio 2021).

References

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.

Dryer, Matthew. 2018. On the order of demonstrative, numeral, adjective, and noun. *Language* 94(4). 798–833.

[4] Consider Smaldino's parable of the cubist chicken for a great explanation of this issue (Smaldino 2017 and https://bjks.buzzsprout.com/1390924/episodes/7048246-8-paul-smaldino-cubist-chickens-formal-models-and-the-psychology-curriculum).

Fried, Eiko I. 2020. Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory* 31(4). 271–288.

Machery, Edouard. 2020. What is replication? *Philosophy of Science* 87(4). 545–567.

McElreath, Richard. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton: CRC Publications.

Muthukrishna, Michael & Joseph Henrich. 2019. A problem in theory. *Nature Human Behaviour* 3. 221–229.

Nichols, John. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.

Pearl, Judea. 2009. *Causality: Models, reasoning and inference*. Cambridge: Cambridge University Press.

Schmidt, Stefan. 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology* 13(2). 90–100.

Seržant, Ilja. 2021. Slavic morphosyntax is primarily determined by its geographic location and contact configuration. *Scando-Slavica* 67(1). 65–90.

Shcherbakova, Olena, Volker Gast, Damián Blasi, Hedvig Skirgård, Russell Gray & Simon Greenhill. 2023. A quantitative global test of the complexity trade-off hypothesis: The case of nominal and verbal grammatical marking. *Linguistics Vanguard* 9(s1). 155–167.

Smaldino, Paul E. 2017. Models are stupid, and we need more of them. In Robin R. Vallacher, Stephen J. Read & Andrzej Nowak (eds.), *Computational social psychology*, 311–331. New York: Routledge.

Szollosi, Aba, David Kellen, Danielle J. Navarro, Richard Schiffrin, Iris van Rooij, Trisha Van Zandt & Chris Donkin. 2020. Is preregistration worthwhile? *Trends in Cognitive Sciences* 24(2). 94–95.

van Rooij, Iris & Giosuè Baggio. 2021. Theory before the test: How to build high-versimilitude explanatory theories in psychological science. *Association for Psychological Science* 16(4). 682–697.

Wimsatt, William C. 2007. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge: Cambridge University Press.