Science
AAAS

Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"

T. Florian Jaeger *et al.*

Science **335**, 1042 (2012);
DOI: 10.1126/science.1215107

Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"

T. Florian Jaeger,[1,2]* Daniel Pontillo,[1] Peter Graff[3]

Atkinson (Reports, 15 April 2011, p. 346) argues that the phonological complexity of languages reflects the loss of phonemic distinctions due to successive founder events during human migration (the serial founder hypothesis). Statistical simulations show that the type I error rate of Atkinson's analysis is hugely inflated. The data at best support only a weak interpretation of the serial founder hypothesis.

Atkinson (*1*) presents evidence for the serial founder hypothesis that the sound structures of languages across the globe reflect a succession of founder events throughout human migration. Based on the hypothesis that "phoneme distinctions are more likely to be lost in small founder populations," Atkinson derives the prediction that the number of different sounds distinguished in a language (its phonemic diversity) should decrease with increasing migration distance to the geographical origin of language.

Because the point of origin of language is unknown, Atkinson pursues the search for the origin and the test of the serial founder hypothesis jointly. For 504 nonextinct languages from (*2*–*5*), he calculates the "total normalized phoneme diversity" as a standardized measure of each language's phonological complexity [see our supporting online material (SOM), section 1]. To test the serial founder hypothesis while simultaneously controlling for a previously documented population size effect (*6*), as well as nonindependence between languages due to genetic relations, he employs linear mixed models. Specifically, phonemic diversity is modeled as a linear combination of (log-transformed) population size, migration distance to the hypothesized point of origin, and the interaction of these two variables, while adjusting for genetic relations between languages (random intercepts for language family, subfamily, and genus) (SOM, section 2).

To determine the most probable point of origin, separate linear mixed regressions are fit for each of 2560 coordinates on the globe corresponding to known locations of extinct or nonextinct languages (*2*). Atkinson finds that the best model fits are obtained for coordinates in southwest Africa. His result is replicated in Fig. 1A, obtained using the same methods as in (*1*) but based on an updated database with 2677 language coordinates (*2*) [see figure 2A in (*1*)]. Atkinson then reports the best-fitting model, which exhibits the predicted negative correlation between phonemic diversity and migration distance to the origin ($P < 0.00003$). This highly significant effect in the predicted direction, together with the clustering of likely origins in southwest Africa, seems to lend strong support for the serial founder hypothesis.

However, as we show below, Atkinson's analysis suffers from a severely inflated type I error rate due to repeated tests on the same data. In three statistical simulations, we find that Atkinson's results only support a weak interpretation of the serial founder hypothesis.

Simulation 1 (SOM, section 4.1) randomly reassigned the 504 languages to language coordinates within language families (the top-level grouping structure reflecting genetic relationships between languages). For example, languages from the Niger-Congo family were randomly reassigned to coordinates of languages from the Niger-Congo family. For each of 10,000 simulation samples, we determined the best fit out of the 2677 possible origins, following the same procedure applied to the original data by Atkinson. The predicted significant negative effect of migration distance on phonemic diversity is found in 100% of the 10,000 random simulation samples. Even effects as strong as those observed by Atkinson are found in 9.8% of our simulation samples. As in the original analysis, all best-fit origins lie in southwest Africa (Fig. 1B). Moreover, for each individual fit, the same clustering of likely origins as in the original data is observed (fig. S3). This shows that, contrary to a literal interpretation of the serial founder hypothesis, the effect observed by Atkinson is independent of the distance of individual languages to the origin. As long as the centers of language families order geographically as in the original sample, the predicted effect is obtained. This result is, however, compatible with a weaker hypothesis: The effect of repeated founding events might only be detectable if aggregated over long periods of time. In that case, the effect is only expected to be visible at the family level (i.e., on data aggregated over language families), which is observed in simulation 1.

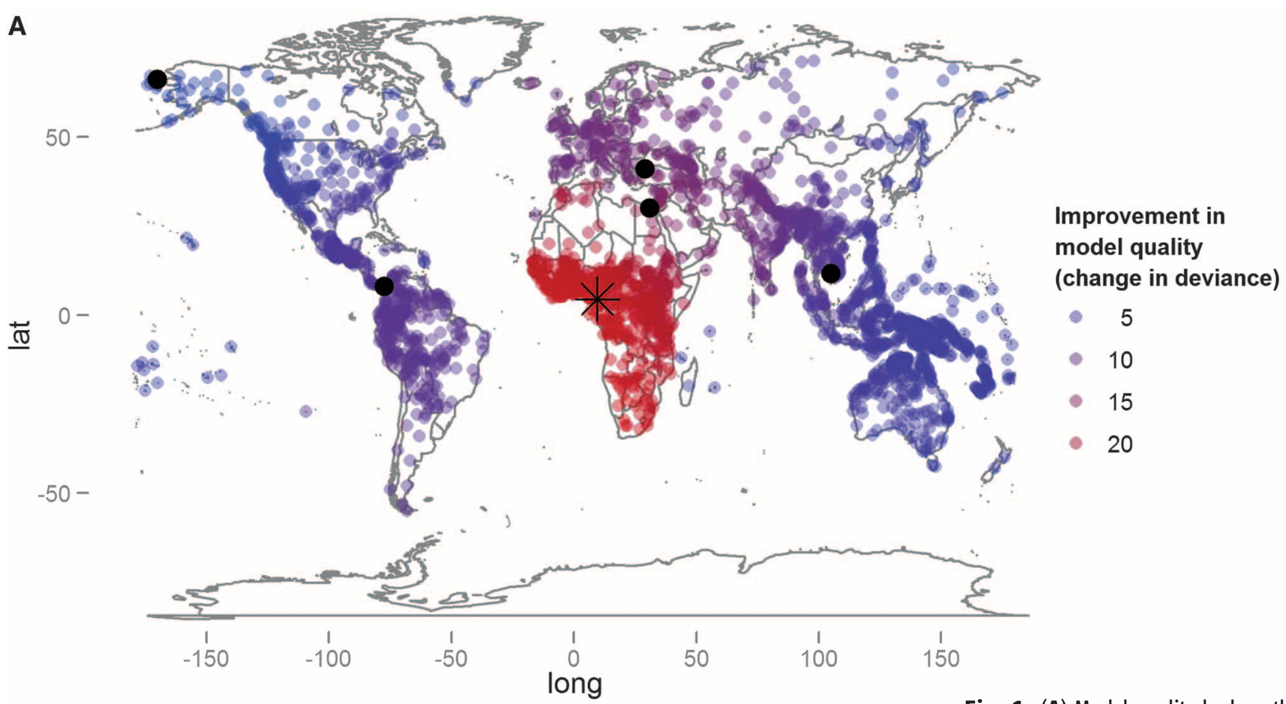
To assess how likely it is that the effect at the family level is due to chance, simulation 2 (SOM, section 4.2) employed a hierarchical sampling procedure: For each simulation sample, language families were randomly reassigned to coordinates around the globe, and languages were randomly reassigned conditionally on the geographical center of their language family (languages in the same family have a strong tendency to cluster geographically). We found the effect of migration distance in the predicted direction for 20.7% of 10,000 samples. This type I error rate is substantially higher than the conventionally accepted 5%. Unsurprisingly, the most probable origins in simulation 2 were more uniformly distributed around the globe than in simulation 1, although, interestingly, the highest proportion of likely origins was again found in Africa (Fig. 1C). This suggests that coordinates in Africa are a priori more likely to be associated with better fits, not because of the origin and direction of human migration but because of any or all of the following: (i) properties of the *World Atlas of Language Structures* data employed by Atkinson (*2*–*5*), such as the distribution of language coordinates across the globe; (ii) the fact that genetically related languages (which tend to share linguistic properties, including phonemic diversity) tend to cluster geographically; (iii) the geography of the globe; and (iv) the constraint that intercontinental migration routes have to pass through the five waypoints shown in Fig. 1A. Simulation 3 (SOM, section 4.3) further suggests that geographic clustering of most probable origins is obtained even in the absence of (ii), although not necessarily in Africa.

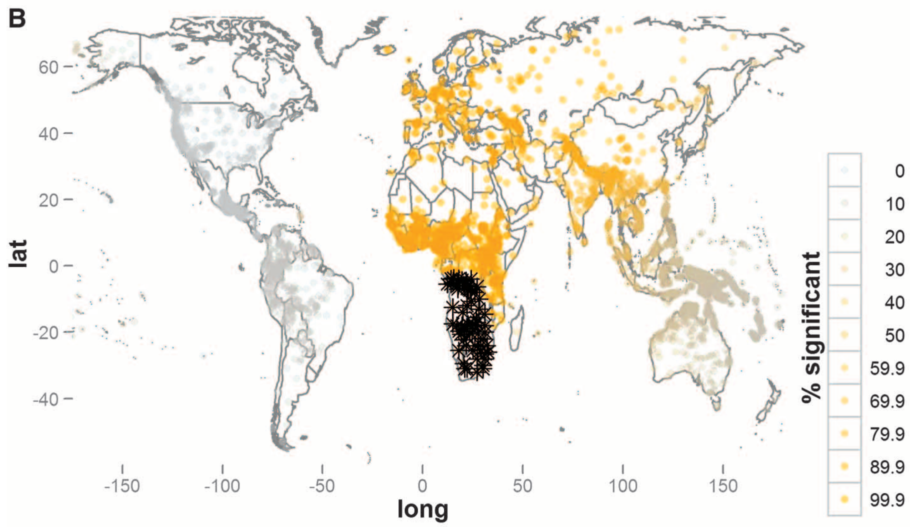
Despite the inflated type I error rates, simulation 2 finds support for the weak interpretation of the serial founder hypothesis: Effect sizes as large as or larger than those reported by Atkinson are observed in only 0.11% of all simulation samples. Although the resulting estimate for an adjusted significance level of the distance effect is two orders of magnitude larger (less significant) than the P value reported by Atkinson, it is still significant ($P < 0.002$). As detailed in the SOM (section 5), this estimate should be taken to be a lower bound (i.e., a best-case scenario for Atkinson).

In conclusion, the literal interpretation of Atkinson's serial founder hypothesis is not supported by Atkinson's data. The data are, however, compatible with a weaker interpretation of the serial founder hypothesis, although it is unclear whether the effect would remain significant under more realistic simulations. There is, nonetheless, reason to be optimistic about the prospect of future evaluations of the serial founder hypothesis: The approach taken here could be used to determine what type of data would be needed

[1]Brain and Cognitive Science, University of Rochester, Meliora Hall, Box 270268, Rochester, NY 14627–0268, USA. [2]Computer Science, University of Rochester, Box 270226, Rochester, NY 14627–0226, USA. [3]Department of Linguistics and Philosophy, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building 32-D808, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: fjaeger@bcs.rochester.edu





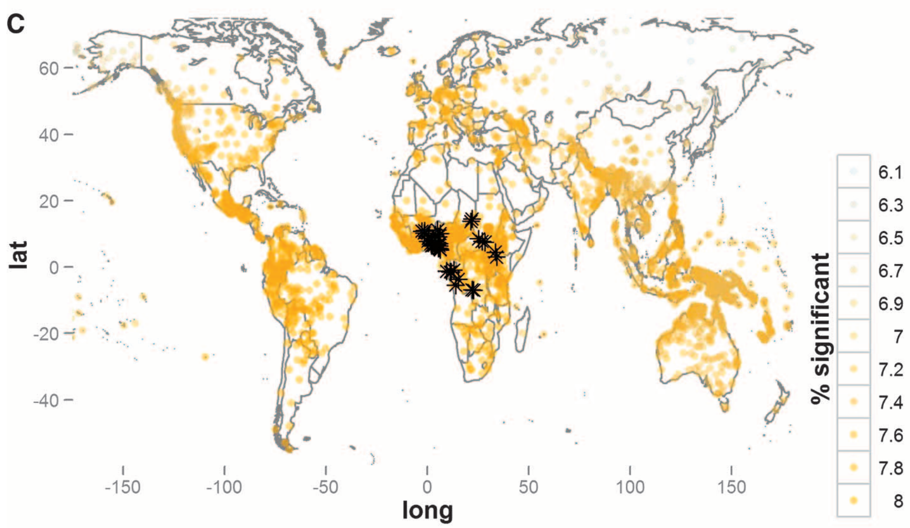


Fig. 1. (**A**) Model quality by hypothesized point of language origin. Colors indicate model quality in terms of the model's deviance compared to a model without the distance predictor. The best fit is indicated by an asterisk. The solid black dots mark the five waypoints that intercontinent migration routes are assumed to pass through [see table S4 in (*1*)]. (**B**) Illustration of the distribution of likely origins found in simulation 1. Points correspond to possible points of origin. Color indicates the percentage of simulation samples for which the hypothesis that the coordinates of the point are the point of origin yielded a significant effect of migration distance. The coordinates for which significant effects were observed most frequently are indicated by asterisks (all within 0.1% of each other). (**C**) Illustration of distribution of likely origins found in simulation 2 (see SOM for details on simulations).

to convincingly test the serial founder hypothesis. For example, simulations could determine how many language families with how many languages are required. Similarly, simulations could shed light on the question from which geographical regions more language data are most urgently needed to answer questions about the origin of language.

References and Notes

1. Q. D. Atkinson, *Science* **332**, 346 (2011).
2. M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, *The World Atlas of Language Structures Online* (Max Planck Digital Library, Munich, 2008).
3. I. Maddieson, in *The World Atlas of Language Structures Online*, M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Max Planck Digital Library, Munich, 2008), pp. 14–17.
4. I. Maddieson, in *The World Atlas of Language Structures Online*, M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Max Planck Digital Library, Munich, 2008), pp. 10–13.
5. I. Maddieson, in *The World Atlas of Language Structures Online*, M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds. (Max Planck Digital Library, Munich, 2008), pp. 58–61.
6. J. Hay, L. Bauer, *Language* **83**, 388 (2007).

Acknowledgments: We are very grateful to Q. D. Atkinson for generously sharing his data with us and answering our questions about the analyses he conducted. The work presented here was partially supported by an Alfred P. Sloan Research Fellowship and the University of Rochester's Wilmot Award to T.F.J.

Supporting Online Material

www.sciencemag.org/cgi/content/full/335/6072/1042-a/DC1
Materials and Methods
SOM Text
Figs. S1 to S9
Table S1
References

11 October 2011; accepted 23 January 2012
10.1126/science.1215107