Appendix: Response to Hammarström

Caleb Everett, Damian Blasi & Sean Roberts

The problem faced by the original paper was how to test a uni-directional effect. We wanted to test whether there is a 'gap' in the distribution of complex tone languages in dry regions (what HHC calls the 'corner' hypothesis). Hammarström's comments (henceforth HHC) criticise several points in our initial study. Here we focus on a few of these criticisms, but it's clear that the following is true of the original paper:

There is no control for area and language family at the same time.
The baseline comparison was not described in enough detail to meet standards of replicability.
The within-continent tests did not control for historical relations, and the correlations may therefore be inflated.

HHC focusses on three different tests. The first test addressed the use of categorical tone language types, while the last two focussed on testing the corner hypothesis.

Test 1: A linear correlation between raw number of tones and humidity, sampling only independent languages (report mean correlation coefficient and mean p value).
Test 2: Measure the difference in percentiles between complex and non-complex tone (the main test in the original paper).
Test 3: Compare the magnitude of these differences in percentiles to a random baseline (the result in the paper referring to the 95% confidence intervals).

We replicated test 1, selecting languages from: independent families as defined in the ANU phonotactics database; independent families as defined in Glottolog[2]; and independent geographic regions according to Autotyp. As in the results of HHC, we found no correlation between the raw number of tones and humidity, looking at the whole range (ANU: $r = 0.03$, $p = 0.55$; Glottolog: $r = 0.001$, $p = 0.63$; Autotyp areas: $r = 0.16$, $p = 0.44$). This also held when sampling languages that were independent in both family and area (ANU: $r = 0.16$, $p = 0.45$; Glottolog: $r = 0.16$, $p = 0.43$) However, if we look at only the lowest 3rd of the humidity range (the range for which we make the 'gap' prediction), then we do see a positive correlation when controlling for language family (ANU: $r = 0.27$, $p = 0.02$; Glottolog: $r = 0.32$, $p = 0.002$), although not for area (Autotyp areas: $r = 0.23$, $p = 0.34$), nor when controlling for both area and family (ANU: $r = 0.28$, $p = 0.29$; Glottolog: $r = 0.28$, $p = 0.28$).

In test 2, HHC criticises the original paper for interpreting a difference in 89% of independent samples as significant, instead of adhering to the more conventional 95% criterion. This rather curious observation seems to stem on the fact that, unfortunately, the target *statistic* can be interpreted as a probability, which might be misleading on a superficial reading. The

[2] Information from Glottolog was linked (by Hammarström) to the data through iso codes from the ANU phonotactics database. However, we discovered several possible errors in the mappings between iso codes and ANU languages while going through the data. The majority of these don't change which language family a language is assigned to but we warn that care should be taken when interpreting these results.

null hypothesis should specify the distribution of the *statistic* - in the null case, what would be the value of the given percentile? *That* is the quantity on which one could impose the conventional alpha=0.05. The fact that the statistic is 88% -or 1%, or 50% or any other positive value- is irrelevant for its significance.

In test 3, HHC gives one interpretation for the null case / baseline (though others are possible): Sample non-complex tone languages as before, but sample a 'complex' group by picking 1 language per family, disregarding tone status. This tests whether there is something 'special' about the complex tone group. In the original paper, the baseline was constructed by permuting complex and non-complex status randomly between languages, then running test 2 with this new data. This tests how likely it would be to get a similar result with a randomly distributed typological feature. We acknowledge that the approach in HHC is more suited to addressing the 'corner' hypothesis, since it randomises only the complex group.

We replicated the results in HHC (in R and Python, results reported here are from Python[3], see SI), taking 5000 samples in each test. Table 1 shows the results. For example, when using glottolog families, complex languages had a higher 15th humidity percentile than non-complex languages in 98.8% of independent samples, and the size of the difference was bigger than 97.7% of samples from the baseline. The results for the ANU families and the 25th percentile are less significant.

		15th	25th	50th	75th
ANU families	Test 2	0.882	0.8598	0.4138	0.4696
	Test 3	0.8438	0.8514	0.372	0.44
Glottolog families	Test 2	0.988	0.8604	0.0346	0.1334
	Test 3	0.977	0.818	0.0368	0.1126
Autotyp geographic areas	Test 2	0.8936	0.9606	0.8656	0.6824
	Test 3	0.8814	0.9614	0.8684	0.6974
ANU families and areas	Test 2	0.9298	0.962	0.8534	0.7312
	Test 3	0.9162	0.9644	0.8718	0.7426
Glottolog families and areas	Test 2	0.806	0.96	0.8536	0.676
	Test 3	0.8064	0.964	0.8656	0.6986

Table 1: Results for different statistical tests.

[3] There are small differences between the results of the two implementations, probably due to the very large number of possible independent samples compared to the number actually taken (5000).

Also included in table 1 are the results for sampling from independent geographic areas (according to Autotyp linguistic areas, which are designed to capture known areas of linguistic contact), and the results are similar.

The final set of results control for family and area at the same time. That is, in each group, 16 languages are selected which do not share a language family nor a geographic area. This test should be more conservative than controlling for just one aspect, so it is surprising that the results are stronger in some cases (not weaker, as predicted in HHC). Given the very small number of languages in each sample, the 15th percentile is a very coarse measure, but we observe similar values when looking at the differences in means across the whole range, instead of percentiles (ANU: test 2 = 0.953, test 3 = 0.968; Glottolog: test 2 = 0.944, test 3 = 0.931, see table 2 below for full results).

Returning to another central point in HHC, should we be testing for a trade-off across the whole scale, or is there a 'corner' effect? While there is not room to resolve that here, and while synchronic analyses may not be the most productive, we make one general point, and then reexamine the data.

Contrary to what is suggested in HCC, the comparison of both distributions at the mean or median value does not bear any relevance for the testing of the 'corner' hypothesis. Any possible direction of the inequality between medians is agnostic with respect to the behaviour in the tail of the distribution. As for the mean - it summarizes information about the whole distribution, so it is not possible in principle to deduce anything about the tail behaviour given such a summary statistic. The percentile measure is a better direct measure of a 'gap', since it only measures differences in the tail of the distribution, and is agnostic to properties of the distribution at higher ranges.

Going back to the data, in figure 2 of the original paper there is a big difference between the distributions of complex and non-complex languages at the lower range, but a much smaller difference at the higher range. However, this cumulative distribution curve represents all languages, including historically related ones, and is therefore misleading. In Figure 1 below, we try to visualise the distribution of languages while controlling for relatedness[4]. Both interpretations may be evident. In the shared area on the left of the graph, there appears to be a trade-off (complex languages are less frequent in low humidity, but more frequent in medium humidity). However, the two distributions look more similar in the area of the graph to the right, suggesting that there is no effect in extremely humid zones.

The results for test 1 when looking at languages in the driest 3rd of the distribution provides some statistical support for this. However, given the other commentaries in this issue, it is not clear whether this is indicative of a difference due to climate or some historical process

[4] The figure shows two density curves, where the height of the curve is the mean density of many independent samples. For each curve, one language from each family is chosen at random and the density curve is estimated for these samples (using the default kernel density estimation function in R). This is repeated many times, then the mean curve height is taken for many points along the x axis. See the SI for the R code, or visit https://github.com/seannyD/ToneClimateJoLE.

affecting specific families in humid regions. Furthermore, not enough is known yet about the precise relationship between humidity and production of tone that would allow a prediction of a particular cut-off. In this case, we simply submit these observations, but suggest that the next steps should focus on the use of more sophisticated, diachronic tests and more nuanced linguistic measures.

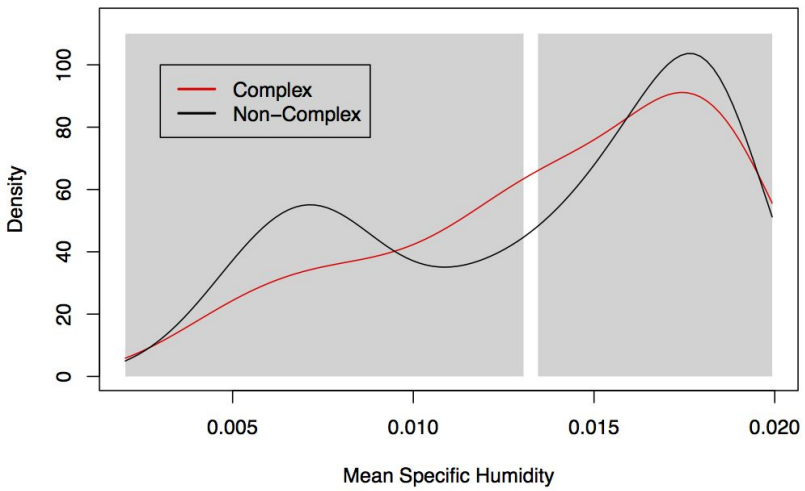

Figure 1: Mean density curve for independent samples for languages with (red) and without (black) complex tone languages. The shaded area on the left covers the first third of languages by humidity, while the shared area on the right covers the remaining two thirds.

Perhaps more important than the results reported here is the general message for large-scale cross-cultural statistical analyses. These tests are complex, and replication in order to confirm results is difficult without explicitly stating the measures and procedures used (or providing code). We acknowledge that the original paper did not do enough to allow replication of our results.

To summarise, not all results reach 95%, but not all tests statistics are necessarily interpretable as p-values. The comparisons against the baselines are a good step in the right direction, but provide mixed results. The result closest to significance comes from applying the most conservative test requested in HHC - controlling for area and family, looking for a difference in means across the whole range ($p = 0.03 - 0.07$). This is without the control for multiple testing, but the tests were designed to answer a single question, so it's not clear that controlling for multiple testing (nor the strict bonferroni approach) is appropriate.

Taking the strictest approach to the results, we cannot rule out the null hypothesis that humidity is correlated with tone. Yet the results remain generally consistent with our hypothesis and therefore suggestive. In general, we would like to argue that there is not a single, 'best' statistical approach to testing any given hypothesis, and that statistical tools provide insights into data, rather than simple answers about the truth or falsity of a claim. Given the complexity of these issues, we suspect that hypothesis will be more carefully

explorable via experimental work and via, some day, the inspection of phonetic corpora that allow for more nuanced examinations of the relevant patterns of pitch/phonation reliance.

One note about the point about the small number of areas tested in the original paper: As we noted then,t the prediction does not apply to all families. The families we report on are the families which have the best balance between number of languages, variation in tone types and variation in humidity. HHC suggests looking at Indo-European or Trans New Guinea, but such a suggestion misses the aim of our study. After all, Indo-European has very few tone languages and trans new guinea only has languages in the most humid 50% of humidity (for which we make no prediction).

Control	**Test 2**	**Test 3**
Glottolog families	0.5812	0.5014
ANU families	0.7696	0.7164
Autotyp areas	0.9542	0.957
Glottolog families and areas	0.944	0.931
ANU families and areas	0.953	0.968

Table 2: Tests 2 and 3 carried out, but calculating the difference in means, rather than percentiles.

Procedure for Test 2 and Test 3

1) For all language families with complex tone, sample 1 complex tone language per family, call this C
2) Sample 1 language per family, from all families, call this R
3) For all language families with non-complex tone, sample 1 non-complex tone language per family, call this N
4) Make sure the groups are the same size (randomly throw out items from the larger groups)
5) Work out the humidity percentile for each group (C, N and R)
6) Return the difference in percentiles: C - N and C - R

Repeat steps 1-6 for 5000 times.
Test 2: Test the proportion of times (C - N) > 0
Test 3: Test the proportion of times (C - N) > (C - R)