Appendix A: Power Analysis

The sample size for our study was determined by conducting a power analysis for both a linear and a polynomial effect of Structure Score on binary accuracy (1 for accurate, 0 for inaccurate). We based our power simulations for this analysis on scaled down effect sizes that were estimated on data collected in a pilot study. Additionally, we computed power for two hypothesized effects of Group Size on binary accuracy.

Here we explain the rationale and procedure of the power analysis, as well as the resulting plots that we used to determine our study's sample size. The simulations and analyses was performed in R (v. 3.5.0; R Development Core Team, 2008) using the *simr* package (v. 1.0.5; Green & McLeod, 2016). The full script for running the simulations and power analyses is available at https://osf.io/abvcg/, and includes detailed comments describing the procedure. The full output of our simulation run is available at https://osf.io/htvqy/.

Effect of structure

Our hypothesized effects of structure on binary accuracy were based on results from a pilot study, in which we tested three out of our ten input languages (S1, S3 and S5), with two participants per language. We expected that the effect of Structure on binary accuracy would either be a linear or a 2-degree polynomial effect, so two separate generalized linear mixed effect models were fitted to the pilot data accordingly.

Figure 1 visualizes the data obtained in the pilot experiment and the estimates of the two models for the fixed effects of Structure. We used these models to simulate the data for the rest of the analyses after scaling down the effect sizes by factors of 0.1, 0.15 and 0.2. **Table 1** lists the effects as estimated by the pilot model, as well as the scaled down effect sizes for each scaling factor.

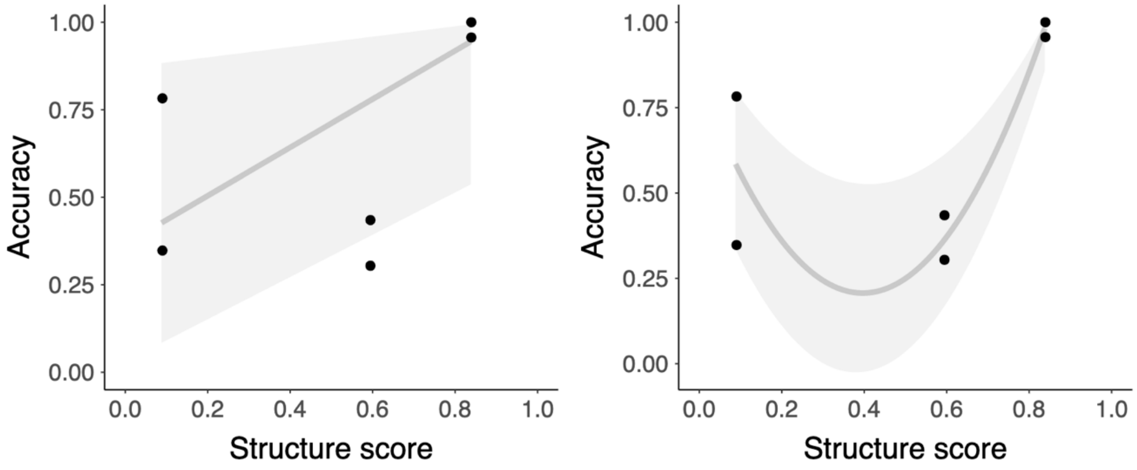

Figure 1. Linear and 2-degree polynomial effects of Structure on binary accuracy, as predicted by the linear mixed-effect models fitted on the pilot dat**a.** Mean accuracy scores per participant are visualized as black dots**.**

Table 1. Model estimates for the fixed effect of Structure on binary accuracy as fitted on the pilot data, as well as the scaled down effects used for our power simulations, by the three different scaling factors (0.1, 0.15 and 0.2).

	Pilot model estimate	0.1-scaled effect	0.15-scaled effect	0.2-scaled effect
Linear model	13.4	1.34	2.01	2.68
Polynomial model, linear term	14.6	1.46	2.19	2.92
Polynomial model, quadratic term	18.5	1.85	2.78	3.70

Effect of group size

We expected the potential effect of group size to be positive, i.e., that participants learning languages created by bigger groups would obtain higher accuracies on the memory test. To simulate an effect of Group Size, we scaled the predicted mean accuracy of the big group languages by a factor of either 1.05 or 1.10, simulating an effect of either 5% or 10% increased accuracy respectively. We chose these effect sizes as we estimated them to be the smallest possible effects on accuracy that would still be reasonably measurable with our planned experimental setup.

Simulation procedure

Power for the effect of group size was calculated for sample sizes ranging from 2 participants to 15 participants per each input language condition (i.e. from 20 to 150 participants in total), and included all combinations of the scaled effects of structure and group size. For each of these 84 possible settings, the simulation was run 1000 times to calculate the rates of correctly detecting each specified effect using linear mixed effect models that were equivalent to the confirmatory analysis for binary accuracy used for our experimental data.

In the case of structure, power was estimated for both detecting an effect and preferring it over the other effect type in model comparison (e.g. correctly detecting a polynomial effect on data simulated from a polynomial model, and preferring the 2-degree polynomial effect over the linear effect). To estimate power for the effect of group size, we calculated the average rate of correctly detecting an effect of group size across all effect types and sizes of structure.

Power simulation results

Our obtained estimates of statistical power varied per simulation setting (i.e., sample size, effect size, and effect type). Below we visualize the power curves for the two different effect types of Structure, as well as the effect of Group Size, for different effect sizes and sample sizes. The script for reproducing these graphs is available at https://osf.io/ywat5/, and the full results of these simulations are available at https://osf.io/htvqy/. Overall, the results show that power varied according to effect size, but rapidly increased when the effect size was higher than our smallest simulated effect size (scaling of 0.1).

Figure 2 visualizes the statistical power for finding a significant effect of Structure on binary accuracy and for preferring the model with the linear fixed effect over the model with the polynomial fixed effect (in both cases with α = 0.05) given that the data was generated by a linear model. **Figure 3** visualizes the statistical power for finding a significant effect of Structure on binary accuracy and for preferring the model with the polynomial fixed effect over the model with the linear fixed effect (in both cases with α = 0.05) given that the data was generated by a polynomial model. **Figure 4** visualizes the statistical power for finding a significant effect of Group Size on binary accuracy (α = 0.05) given that the data was generated with an either 5% or 10% increase in accuracy for bigger groups.

Based on these results, we decided on a sample size of 10 participants per input language condition, corresponding to 100 participants in total. This sample size provided us with reasonable statistical power (>60%) even for very small effect sizes. Importantly, the effect sizes of Structure on binary accuracy estimated by the models fitted on our *actual* experimental data were higher than our largest simulated effects (scaling of 0.2). As such, we are confident that the statistical power for our confirmatory analysis was over 80%.

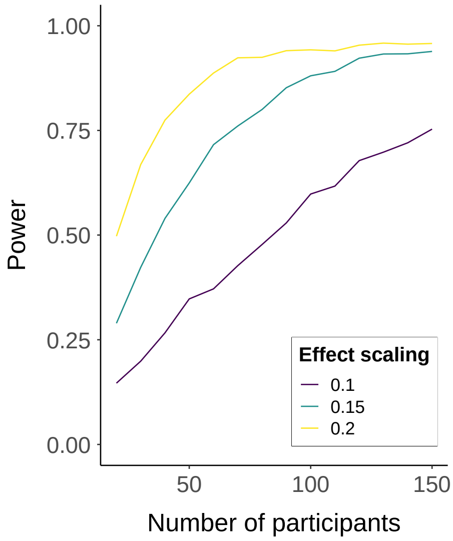

Figure 1. Power curves as estimated based on our simulations for a linear effect of Structure on binary accuracy. Shadings indicate the 95% binomial confidence intervals. Depending on effect size the power for our chosen sample size of 100 participants is approximately between 60% and 90%.

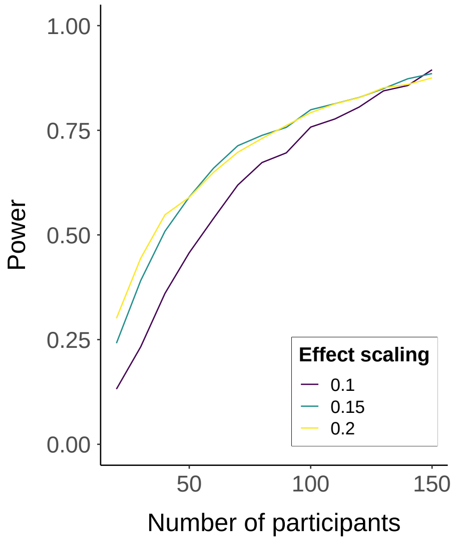

Figure 2. Power curves as estimated based on our simulations for a second-degree polynomial effect of Structure on binary accuracy. Shadings indicate the 95% binomial confidence intervals. Depending on effect size the power for our chosen sample size of 100 participants is approximately between 70% and 80%.

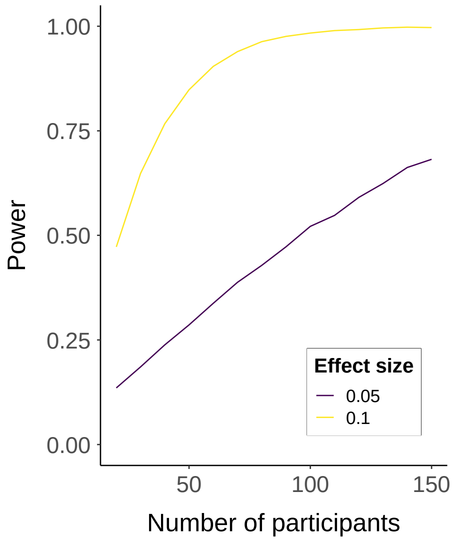

Figure 3. Power curves as estimated based on our simulations for 5% vs. 10% effect of Group Size. Shadings indicate the 95% binomial confidence intervals. Depending on effect size the power for our chosen sample size of 100 participants is over 60%.

References

Green, P., MacLeod, C.J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, **7**(4), 493–498. doi: 10.1111/2041-210X.12504, https://CRAN.R-project.org/package=simr.

R Development Core Team (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.