**Appendix B: Comparisons with the Short Version**

In addition to the 24 groups reported in the paper, we also collected data from six small groups of four participants who played a shorter version of the experiment, which included only eight rounds instead of 16. These groups were tested in the first batch of data collection. The report of the individual performance of these six groups can also be found in Raviv, Meyer & Lev-Ari (2019).

In this appendix, we report the results of the comparison between these small “short-version” groups and the twelve larger groups reported in the paper. We compared the performance of these groups twice: first during the first eight rounds, when groups have the same amount of exposure but differ in their shared history; and then again at the end of the experiment (i.e., the seventh and eighth round for the small groups vs. the 15th and 16th round for the larger groups). At that point, the amount of shared history is equated - members of both types of groups have interacted with each other twice by that time point – but the amount of exposure differs.

**Participants**

A total of 24 adults participated in the short version (mean age=23.2y, SD=4.53; 18 women). All participants were native Dutch speakers and were recruited using the participant database of the Max Planck Institute for Psycholinguistics. Participants in the short version were paid 20€ or more depending on the time they spent in the lab (between 120 to 150 minutes).

**Stimuli and Procedure**

The stimuli and procedure of the short version were identical to those reported in the main paper, except for the fact that the experiment ended after the first eight rounds (i.e., participants completed the test round at round 8, filled the debriefing form and then left, without having lunch and without reconvening to continue the second half).

**Analyses**

We used mixed-effects regression models to test the effect of community size on all measures, using two types of models: (1) Models that analyze changes over the course of the first eight rounds; (2) Models that compare the final languages created by small and larger groups before and after the additional rounds, that is, at rounds 7 and 8, and at rounds 15 and 16. All models were generated using the lme4 and pbkrtest packages in R (see references in the main paper). All the data and the scripts for generating the models can be found online at <https://osf.io/y7d6m/>.

Models of type (1) were used to predict changes in the dependent variable as a function of time and community size, and included all data from thesix small groups who played the short version and data the first eight roundsof the 12 larger groups who played the full version and were collected in the same batch. Models for communicative success included data from communication rounds only (excluding the eighth test round). In models for communicative success, convergence, and stability, the fixed effects were condition (dummy-coded with small group as the reference level), round number (centered), item current age (centered), and the interaction terms Condition X item current age and condition X round number. The random effects structure of models for communicative success, convergence, and stability always included by-scenes and by-groups random intercepts, as well as by-scenes and by-groups random slopes with respect to the effect of round number. Because the structure score was calculated for each producer over all scenes in a given round, the model for linguistic structure did not include item current age as a fixed effect. The model for linguistic structure therefore included fixed effects for round number (linear and quadratic terms, centered), condition (dummy-coded with small group as the reference level), and the interaction term condition X round number (linear and quadratic terms). The model for linguistic structure included random intrecpects and random slopes for the effect of round number (linear and quadratic terms) with respect to different producers who were nested in different groups.

Models of type (2) were used to compare the final languages created by small and larger groups in rounds 7-8 and in rounds 15-16, whenever we found evidence for group size influences in type (1) models. Since group size did not influence communicative success, we did not run models of type (2) for this measure. The fixed effect was a three-level categorical variable (i.e., small groups at rounds 7-8, larger groups at rounds 7-8, larger groups at rounds 15-16). This variable was dummy-coded with small groups at rounds 7-8 as the reference level. In models for convergence and stability, the random effects structure included random intercepts for different groups and different scenes. In models for linguistic structure, the random effect structure included random intercepts for different producers nested in different groups.

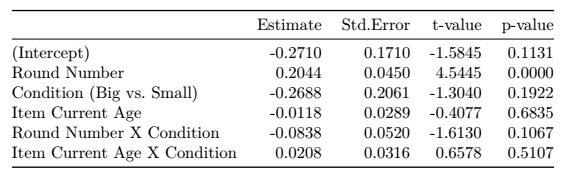
**Results**

*Communicative Success*

Communicative Success increased during the first eight rounds (Model 1: β=0.20, SE=0.05, t=4.5, p<0.0001). Participants became more accurate as rounds progressed, and this increase was not affected by community size (Model 1: β=-0.27, SE=0.2, t=-1.3, p=0.19). While small groups were more accurate than larger groups at the seventh round (Model 2: β=-0.11, SE=0.03, t=-3.86, p=0.0009), larger groups reached more accuracy when given additional rounds (Model 2: β=0.11, SE=0.03, t=3.87, p=0.009).

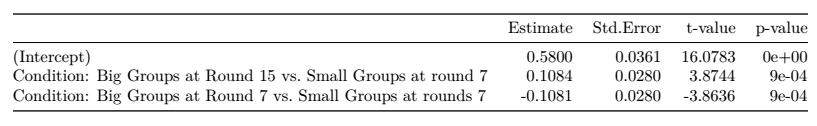
1. *Type (I) Model: Accuracy during the first 8 rounds*

Accuracy ~ centered.Round \* Condition + centered.ItemCurrentAge \* Condition + (1 + centered.Round | ItemID) + (1 +centered.Round | Group)



1. *Type (II) Model: Final Accuracy comparison at round 7 and round 15*

MeanAccuracy ~ Condition + (1 | Group) + (1| ItemID)

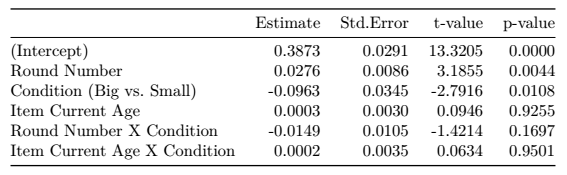


*Convergence*

Convergence increased significantly during the first eight rounds (Model 3: β=0.03, SE=0.01, t=3.2, p=0.004), with participants in the same community aligning over time and using more similar labels. Larger groups were significantly less converged than small groups during the first eight rounds (Model 3: β=-0.1, SE=0.03, t=-2.79, p=0.01). A comparison of convergence levels before and after the additional rounds confirmed that larger groups were significantly less converged by the end of the eighth round (Model 4: β=-0.08, SE=0.02, t=-4.11, p=0.0005). However, this disadvantage disappeared once larger groups completed all 16 rounds and had the same shared history (Model 4: β=0.02, SE=0.02, t=1.2, p=0.245). This result suggests that larger groups needed more time in order to develop globally shared languages, but eventually reach similar levels of convergence as small groups.

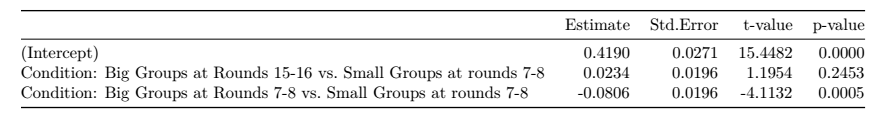
1. *Type (I) Model: Convergence during the first 8 rounds*

Convergence ~ centered.Round \* Condition + centered.ItemCurrentAge \* Condition + (1 + centered.Round | ItemID) + (1 +centered.Round | Group)



1. *Type (II) Model: Final Convergence comparison at rounds 7-8 and rounds 15-16*

MeanConvergence ~ Condition + (1 | Group) + (1| ItemID)

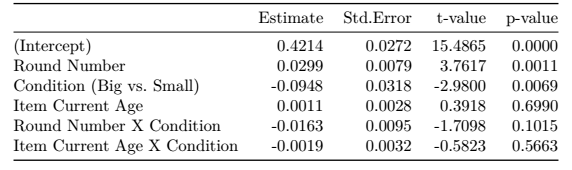


*Stability*

Stability significantly increased during the first eight rounds, with participants using labels more consistently over time (Model 5: β=0.03, SE=0.01, t=3.76, p=0.0011). Larger groups were significantly less stable than small groups during the first eight rounds (Model 5: β=-0.09, SE=0.03, t=-2.98, p=0.0069), and a comparison of the stability levels before and after the additional rounds confirmed that by the end of the eighth round, larger groups showed less stability compared to small groups (Model 6: β=-0.08, SE=0.02, t=-5.3, p<0.0001). Yet again, this pattern disappeared once larger groups were given the additional rounds (Model 6: β=0.025, SE=0.02, t=1.55, p=0.134). That is, while larger groups needed more time to develop consistent languages, they eventually reached the same level of stability as small groups.

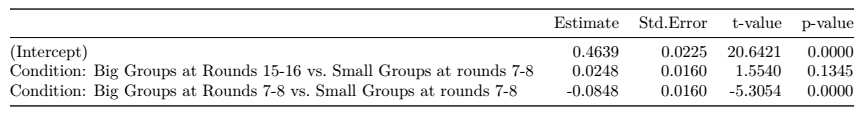
1. *Type (I) Model: Stability during the first 8 rounds*

Stability ~ centered.Round \* Condition + centered.ItemCurrentAge \* Condition + (1 + centered.Round | ItemID) + (1 +centered.Round | Group)



1. *Type (II) Model: Final Stability comparison at rounds 7-8 and rounds 15-16*

MeanStability ~ Condition + (1 | Group) + (1| ItemID)



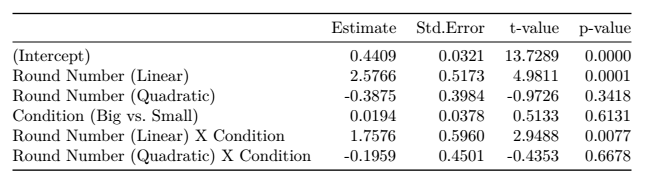
*Linguistic Structure*

Linguistic Structure significantly increased over the first eight rounds in a linear way (Model 7: β=2.58, SE=0.52, t=4.98, p=0.0001), with participants’ languages becoming more systematic over time. This increase in structure was modulated by group size (Model 7: β=0.02, SE=0.04, t=2.95, p=0.0077), with participants in larger groups developing structured languages faster compared to participants in small groups. Although the languages of small and larger groups were equally structured after eight rounds (Model 8: β=0.0005, SE=0.01, t=0.36, p=0.97), members of larger groups developed languages with significantly more linguistic structure after given additional rounds (Model 8: β=0.066, SE=0.01, t=5.3, p<0.0001).

1. *Type (I) Model: Linguistic Structure over time*

Linguistic Structure ~ poly(centered.Round,2) \* Condition

+ (1 + poly(centeredRound ,2) | Group/Producer)



1. *Type (II) Model: Final Linguistic Structure comparison*

MeanStructure ~ Condition + (1 | Group/Producer)

