Replication of

Experimental Evidence for the Influence of Group Size on Cultural Complexity

by Derex, M. / Beugin, M. P. / Godelle B. / Raymond, M. (2013)
in: Nature, 503, pp. 389–391

Replication Authors:
Teck-Hua Ho and Hang Wu

Derex et al. (2013) randomly assigned participants to groups of 2, 4, 8 and 16 players to play a dual-task computer game. Results showed that cultural evolution strongly depends on population size, as players in larger groups maintained higher cultural complexity. They found that as group size increases, cultural knowledge deteriorates less, improvements to existing cultural traits are more frequent, and cultural trait diversity is maintained more often.

In the dual-task game, players had to collect resources individually to improve their individual 'health'. A cultural package composed of two demonstrations – one concerning a simple task and one concerning a complex task – was introduced within groups of different sizes (2, 4, 8 or 16 players). The players were told that each item in the cultural package could be improved. During each of the 15 trials of the game, each player had to build an arrowhead (simple task) or a fishing net (complex task) to collect 'life units' to increase health (Extended Data Fig. 1). The cultural trait diversity of the group thus consisted of some players building one artifact, while the remaining players built the other; diversity was lost when all individuals built the same object. In our replication, we will include all group sizes used in the original study.

Hypothesis to replicate and bet on:

The probability of maintaining cultural diversity (that is, observing both tasks in the group) increases with group size; $\chi^2(1) = 16.3$, the p-value < 0.0001 (exact 0.000054) (p. 389; measured at the group level with group sizes, 2, 4, 8, and 16).

Power Analysis and Criteria for Replication: First Data Collection

The original sample size was 51 groups with a total of 366 male individuals; the standardized effect size measured as the correlation coefficient (r) was 0.525. To have 90% power to detect 75% of the original effect size, 65 groups (17 groups of 2 players, 16 groups each of 4,8, and 16 players) with a total of 482 male individuals are required. The criteria for replication is an effect in the same direction as the original study and a p-value < 0.05 (χ^2 test).

Power Analysis and Criteria for Replication: Second Data Collection

If the original result is not replicated in the first data collection a second data collection of

90 additional groups (24 groups of 2 players, 22 groups each of 4,8, and 16 players) with a total of 664 male individuals will be carried out so that the total sample size is 155 groups (1,146 male individuals) If a second data collection is carried out, we will test whether the original result replicates in the pooled sample of the first and second data collections.

To have 90% power to detect 50% of the original effect size, a sample size of 155 groups is required; i.e. a sample size of 90 groups is needed in the second data collection in order to have a total sample size of 155 groups for the pooled first and second data collections. The criteria for replication is an effect in the same direction as the original and a p-value < 0.05 (χ^2 test) in the pooled data.

Sample

The sample size in the first data collection consists of 65 groups (482 male students) from the National University of Singapore (NUS).

If the original result is not replicated in the first data collection (χ^2 test, p-value < 0.05 in the original direction) a second data collection of 90 additional groups (664 male students) from NUS will be carried out so that the total sample size is 155 groups (1,146 male students).

Materials

We use the same computer program (programmed in Object Pascal with Delphi 6) as used in the original article, provided by the original authors.

Procedure

We follow the procedure of the original experiment. Subjects will be recruited through recruiting advertisements posted in the NUS campus, as well as e-mail invitations sent to an existing voluntary database of undergraduate students maintained by the Centre for Behavioural Economics (CBE) at NUS. Each participant will be randomly assigned to one condition of the experiment. The following summary of the experimental procedure is based on the additional methods section of the online version of the paper.

The participants played a computer game (programmed in Object Pascal with Delphi 6) during which they had to maximize their 'health' using two virtual tasks, making an arrowhead or a fishing net. Before the beginning of the game, players were advised that the fishing-net task was potentially more effective than the arrowhead task but that the fishing-net construction was more difficult. The participants were also informed that the performance of an arrowhead depended only on its shape, whereas the performance of the fishing net depended on its shape and the procedure used to build it. Each player began the game by observing a video demonstration of each task from a cultural package and was instructed that the arrowhead and fishing-net demonstrations could be improved. The arrowhead demonstration involved 15 steps and was associated with a score of 1,638. The fishing-net demonstration involved 39 steps (the sequence of which mattered) and was associated with a score of 2,665. The participants were not aware of the highest achievable score for any task.

The players then had 15 trials to collect resources and improve their health score. At each trial, they had the opportunity to build either an arrowhead or fishing net. Players began the game with a health score of 3,400 units. At each trial, their health level was reduced by 1,000 units, corresponding to their daily needs. Between trials, players could benefit from social information (see below).

Construction period. During the construction period (limited to 90 s), the players had to choose between the arrowhead task and the fishing-net task to collect resources.

The arrowhead task. The performance of an arrowhead depended only on its shape. The arrowhead score ranged from 0 to 2,400 units. A simple symmetric, triangular arrowhead constituted an acceptable performance equal to the player's daily needs. As a consequence, the probability of a non-experienced player scoring below his daily needs was low.

Construction details for the arrowhead task. First, the players had to choose the rectangular grid dimension on which to draw the arrowhead (30 possible values, Extended Data Fig. 1.a). Once the grid was chosen, the players had to draw their arrowhead. By clicking on the grid, the players could draw lines between points (Extended Data Fig. 1.b). The players had to draw the outline of their arrowhead and the virtual relief. No construction rules were implemented.

Score calculation for the arrowhead task. Once an arrowhead was drawn, it was evaluated by the program. The arrowhead was scanned pixel by pixel to evaluate five parameters: the size (α) and the symmetry (β) of the arrowhead, the number of notches (γ) and their regularities (δ), and the triangular shape (λ). All the parameters were compared to a theoretical optimal value and normalized from 0 to 1. The score S was then obtained according to this formula:

$$S = \alpha \cdot 400 + \beta \cdot 800 + \gamma \cdot 800 + \delta \cdot 400 + \lambda \cdot 400$$

The fishing-net task. The participants had access to several virtual tools with which to build their nets. The performance of a net depended on its shape and the procedure used to build it. The net's score ranged from 0 to 5,135 units. Departure from the construction rules (which were unknown to the players) resulted in increased penalties during use of the fishing net. As a consequence, the probability of a non-experienced player scoring below his daily needs (1,000 units) was high.

Construction details for the fishing-net task. First, the players had to choose the squared grid dimension on which to build the net (30 possible values, Extended Data Fig. 1c). Once the grid was chosen, the players had access to different types of ropes and knots, as in a previous experiment. A rope could be set between any pair of attaching points, and a knot could be tied to any attaching point, in any order (Extended Data Fig. 1d). There were limited ropes and knots available. Each additional rope placed on the frame decreased the length of the remaining rope according to the length used. This remaining quantity was visible on the screen.

There were three different types of rope available (thick (red), medium (blue) and thin (green)). Each additional knot placed on the net decreased the length of the remaining knot quantity according to the type of knot used (three sizes available). This remaining quantity of knots was visible on the screen. Modification of one parameter produced complex interactions with others to generate a complex fitness landscape. For example, the use of the thickest ropes prevented the net from breaking but increased the net visibility so that the number of potentially caught fish was reduced. In addition, the order of construction (the process), was important. For example, two ropes that intersect at an attaching point should be tied together with a knot before another rope is put on the frame. If this step is omitted, the expected score is reduced.

Score calculation for the fishing-net task. Once a fishing net was constructed, it was evaluated by the program. A global resistance score (GR) was calculated according to the number of knots and compared to the required number. A local resistance score (LR_i) was determined for each mesh i according to the length and thickness of the ropes involved.

During each virtual fishing exercise, 79 fish were launched, with a unique size of 65 (arbitrary units). The probability of each fish encountering the net increased according to the net overall size (set by the grid spacing parameter) and decreased according to its visibility. The visibility of a net was computed as the sum of the lengths of all ropes used, weighted by their thicknesses. Once a fish was set to interact with the net, random coordinates were generated to identify at which mesh the interaction took place. If the fish was smaller than the mesh, it escaped. If it was larger, the probability of the net breaking was calculated as $1 - (GR \cdot LR_i)$. In such a case, the whole fishing process stopped. If the net did not break, the fish could escape with a probability P_esc, which depended on the shape of the mesh and construction-rule penalty. If the fish did not escape, its size was added to the player's score. This process was repeated until the last fish was encountered or until the net broke.

Information period. After each trial, the resulting score, along with the player's health level, was displayed. The players could also see score lists for the arrowheads and fishing nets generated by the player's group members at the previous trial, ordered by performance. During the first three trials, the cultural package (arrowhead or fishing net) was included in the corresponding list.

By clicking on a score, the players could see the step-by-step procedure needed to build the selected item. Any demonstration lasted 40s, regardless of the number of building steps. At each information period, a player could see only one demonstration. From the fourth information period, cultural-package demonstrations were removed from the lists. The players then had access only to their group member's demonstrations. The duration of the social-information period was 70s.

Rewards calculation. The individual rewards were €10.00 on average. Players who died during the game (health level dropped below 0) earned €2.00. The other players earned an amount €A calculated according to this formula:

$$A = \frac{H_p}{H_g}[5N + 3N_d] + 5$$

where H_p is the player's health level, H_g is the sum of the group's health levels, N is the size of the group, and N_d is the number of dead players within the group.

Treatments. Four group sizes were considered: 2 players, 4 players, 8 players and 16 players. All treatments were replicated 12 times, except for the 2-player treatment, which was replicated 15 times.

Analysis

The analysis will be performed exactly as in the original paper, in which the probability of maintaining cultural diversity (that is, observing both tasks in the group) increases with group size ($\chi^2(1) = 16.3$, $p < 0.0001$). The same test will be used in the replication.

The results will first be estimated based on the first data collection. If the original result is replicated in the first data collection (χ^2 test, p-value < 0.05 in the same direction as the original study), the second data collection will not be carried out.

If the original result is not replicated in the first data collection, a second data collection will be carried out. The above statistical test will then be estimated for the pooled sample of the first and second data collections to test if the original result replicated (χ^2 test, p-value < 0.05 in the same direction as the original study).

The experiment will be conducted in English.

Differences from Original Study

The replication procedure is the same as that of the original study, with some unavoidable deviations. The replication will be performed at the Center for Behavioral Economics (CBE) at NUS between September 2016 and September 2017, whereas the original study was conducted at the Laboratory of Experimental Economics of Montpellier (LEEM) in 2013. In the original experiment, the game (tutorial and interface) and the associated database were in French, while the replication experiment will be in English.

In the original study, the participants received €2.00 to €6.00 show-up fee plus a performance-based payment (on avearge €10.00). In the replication, the participants will receive S$5.00 (in Singapore Dollars) show-up fee plus a performance-based payment calculated (on average S$10.00) using the same rewards calculation formula as in the original experiment.

Replication Results for the First Data Collection (90% power to detect 75% of the original effect size)

[To be added when replication experiments have been completed.]

Replication Results for the First and Second Data Collection Pooled (90% power to detect 50% of the original effect size)

[To be added when replication experiments have been completed.]

Unplanned Protocol Deviations

[To be added when replication experiments have been completed.]

Discussion

[To be added when replication experiments have been completed.]

References

Derex, M. / Beugin, M. P. / Godelle, B. / Raymond, M. (2013): *Experimental evidence for the influence of group size on cultural complexity*, Nature, 503, pp. 389–391.