Replication of

Why Testing Improves Memory: Mediator Effectiveness Hypothesis

by Pyc, M. A. / Rawson, K. A. (2010)
in: Science, 330 (6002), p. 335

Replication Authors:
Colin Camerer, Taisuke Imai, Dylan Manfredi, and Gideon Nave

In a laboratory experiment, Pyc and Rawson (2010) presented undergraduate participants with 48 pairs of Swahili-English translation pairs. Participants received instructions about how to generate keywords (mediators that look or sound similar to the foreign language cue and are semantically related to the English target).

In an initial study trial, and each restudy trial, all participants generated and reported a keyword. Participants were randomly assigned to one of two conditions: (1) test-restudy, where each practice trial involved a cued recall test followed immediately by restudy, and (2) restudy only, where each trial involved only restudies. One week later, participants completed a final cued-recall test: they were shown the cue and asked to recall the mediator they had developed during practice, and then recall the target. Retrieval of both mediators and targets at final test was greater with test-restudy practice than with restudy practice.

Note: the study reports several additional treatments that differed in the manner that the retrieval test was conducted.

Hypothesis to replicate and bet on:

Retrieval of mediators is greater with test-restudy practice than with restudy practice; a comparison of mean mediator retrieval between the test-restudy and the restudy treatments within the CMR treatment, p. 335, $t(34) = 2.37$ and p-value $= 0.02$, t-value and p-value from authors). Note that a successful retrieval in each of the final test questions is defined as correctly recalling any of the keyword mediators that had been generated during session 1.

Power Analysis and Criteria for Replication: First Data Collection

The original sample size was $n = 36$ (two treatment groups out of six, in a study with $n = 118$). The standardized effect size measured as the correlation coefficient (r) was 0.377. To have 90% power to detect 75% of the original effect size a sample size of 132 is required. The criteria for replication is an effect in the same direction as the original study and a p-value < 0.05 (in a two-sided 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

174 additional individuals will be carried out so that the total sample size is 306. If a second data collection is carried out, the criteria for replication is an effect in the same direction as the original study and a p-value < 0.05 (in a two-sided test) in the pooled data.

Sample

The sample size in the first data collection consists of 132 participants from the Wharton behavioral lab subject pool. If the original result is not replicated in the first data collection (two-sided p-value < 0.05 in the original direction) a second data collection of 174 additional individuals from the same population will be carried out so that the total sample size is 306.

Materials

We use the same computer program as used in the original article. Materials include 48 Swahili-English translation pairs (e.g., wingu—cloud) previously normed for item difficulty.

Procedure

We follow the procedure of the original article. The following summary of the experimental procedure is therefore based on the procedure described in the Supplementary Information.

We will randomly assign the participants to one of two groups (practice: test-restudy or restudy only). Participants will receive instructions about how to generate keywords (keywords are mediators that look or sound similar to the foreign language cue and are semantically related to the English target). Each item will be presented for an initial study trial and then three blocks of practice trials. In the test-restudy group, each practice trial for an item will involve a cued recall test followed immediately by restudy. In the restudy group, each trial will involve only restudy. On the initial study trial and each restudy trial, all participants will generate and reported a keyword mediator (e.g., for the translation pair "wingu—cloud" a keyword commonly generated by participants was "wing"). All trials will be self-paced. Exactly seven days later, participants will complete a final cued-recall test. They will be shown the cue with a prompt to recall any of the keyword mediators they had developed, and then recall the target.

As in the original study, participants will complete the study in the lab, which is a large room with isolated compute stations. A research assistant will always in the room in order to answer any questions, and to ensure that the room remained quiet throughout the session.

Analysis

The analysis will be performed exactly as in the original article. In the original article an independent samples t-test was used to compare the mean fraction of mediators correctly recalled at the final test between the test-restudy practice group and the restudy practice group (with a successful retrieval in each of the final test questions defined as correctly recalling any of the keyword mediators that had been generated during session 1). In the original study the mean recall of mediators was 34% in the restudy practice group and 51% in the test-restudy practice group; $t(34) = 2.37$ and $p = 0.0236$ (the t-value and p-value was not reported in the original article, but was received from the original authors). 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 (a two-

sided 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 is replicated (a two-sided p-value < 0.05 in the same direction as the original study).

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 Wharton behavioral lab between September 2016 and September 2017, whereas the data in the original study was carried out at Kent State University in Ohio in spring 2009 (beginning in spring semester, with data collection going through the summer and into the fall).

In the original study participants received 4 experimental course credits for participation. The first session lasted around 90 minutes (3 credits) and the second was less than 30 minutes (1 credit). The replication will have a show-up fee of $20 for each of the two sessions.

The original study had 6 treatment groups, 2 practice (test-restudy or restudy only) $\times$ 3 final test format (C, CM, or CMR) design — we will only have the two conditions of the CMR final test format.

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

Pyc, M. A. / Rawson, K. A. (2010): *Why testing improves memory: Mediator effectiveness hypothesis*, Science, 330 (6002), pp. 335–335.