Replication of

Writing About Testing Worries Boosts Exam Performance in the Classroom

by Ramirez, G. / Beilock, S. L. (2011)
in: Science, 331, pp. 211–213

Replication Authors:
Nick Buttrick, Anup Gampa, Lilian Hummer, and Brian Nosek

In a lab study with members of the University of Chicago community, Ramirez and Beilock (2011) randomly-assigned participants to either expressively write about an upcoming high-stakes math test or simply sit quietly and wait. Expressively-writing participants improved their performance relative to a pretest and performed better than quietly-sitting participants (who performed worse than their pretest). The paper included 4 studies. Studies 1 and 2 are lab studies, and studies 3 and 4 are in-classroom field studies. Study 1 is the study being replicated, study 2 added an additional unrelated-to-the-task writing condition, and in study 3 and its replication, study 4, a class of 9^{th} grade biology students were randomly-assigned to either an expressive writing condition or a control in which they were to think about an unrelated topic before taking an end-of-the-year test

Hypothesis to replicate and bet on:

In a high-pressure in-lab math test, those writing for 10 minutes about their deepest thoughts and feelings regarding the upcoming test improve more on that test compared to simply sitting quietly; an F-test, $p < 0.05$ using a two-tailed test.

Original test statistics: $N = 20$ (10 in each condition); Expressive writing $\mathrm{M}_{pre} = 0.86$ ($SD = 0.09$), $\mathrm{M}_{post} = 0.91$ ($SD = 0.05$), Control $\mathrm{M}_{pre} = 0.82$ ($SD = 0.09$), $\mathrm{M}_{post} = 0.70$ ($SD = 0.11$); $F(1, 18) = 30.53$; $p = 0.00003$ (reported as $p < 0.01$, p. S11).

Power Analysis and Criteria for Replication: First Data Collection

The original sample size was 20 observations, 10 in each of two conditions. The effect size measured as an r was 0.793. Following the protocol of this replication project, to have 90% power to detect 75% of the original effect size a sample size of 25 is required. We will recruit a 26^{th} participant so that we can have equal numbers between conditions. The original authors conducted an orthogonal manipulation of pressure that was reported only in their Supplemental Materials. On recommendation of the original authors, we added the orthogonal manipulation to assess whether the necessary pressure was induced in the main study conditions, measured as a difference in felt anxiety between the main and manipulation-check-comparison conditions. The effect size of the original effect of manipulation was $d = 0.99$. Adding another 26 participants for these manipulation-

check-comparison conditions will give us 93% power to detect an effect size equal to the manipulation-check difference. The criteria for replication is a focal-test effect in the main conditions the same direction as the original study and a p-value < 0.05 (two-sided test).

Power Analysis and Criteria for Replication: Second Data Collection

According to the replication project protocol, if the original result is not replicated in the first data collection of the 26 participants in the main conditions, an additional data collection of 40 individuals in the main conditions will be carried out, for a total sample in the main conditions of 66. 40 participants will then additionally be recruited for the manipulation-check-comparison conditions, leading to a total sample size in all conditions of 132. If a second data collection is carried out, it will be tested if the original result replicates in the pooled sample of the participants of the main condition in the first and second data collections.

To have 90% power to detect 50% of the focal effect, a sample of 66 is required for the main study conditions; i.e. a sample size of 26 in the first collection and 40 in the second collection. With a total of 66 participants in the main study condition, and an additional 66 in the manipulation-check comparison conditions we would have 99% power to detect the original manipulation-check size of $d = 0.99$, and 80% power to detect 50% of the manipulation-check effect size. The criteria for replication is a focal effect in the same direction as the original in the main study conditions and a p-value < 0.05 (in a two-sided test) in the pooled data.

Sample

The sample size in the first data collection will consist of 52 individuals from the University of Virginia, 26 in the main conditions and 26 in the manipulation-check-comparison conditions. Participants will be recruited using the UVA research participant pool. The original authors expressed concern that the University of Chicago participants may be higher achieving on average than UVA students. As such, participation will be restricted to students who scored better than 1400 on their SAT or better than 30 on their ACT. Participants will be compensated with research credit. All participants assigned to the high-pressure scenario will receive an additional $10 in earnings, regardless of their performance.

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 80 additional individuals from the population will be carried out, 40 in the main conditions and 40 in the manipulation-check-conditions, so that the total sample size is 132.

Materials

We will use the same modular arithmetic problems; state-form of the STAI; Expressivity scale; writing condition prompt; High-pressure scenario protocol; and Low-pressure scenario protocol as the original study, as described on pages 3–4 of the Supplementary Information. The experiment will be in English as in the original study.

Procedure

We follow the procedure described in the original article. The following summary of the experimental procedure is based on pages 211–212 of the main article and pages 2–9 of the Supplementary Information as well as from direct feedback provided by the original authors.

Participants will complete the study individually. After giving informed consent, they

will receive background about the materials of the study, and have 8 practice modular arithmetic problems to ensure they understand the task. After the initial 8 problems, they will receive 40 more on a computer (half Low-Demand, half High-Demand), in what appear to be a continuation of the practice session (but in reality make up the "pretest" period of the study). Each problem will begin with a 500ms fixation cross. After the participant answers, there will be feedback denoting either "correct" or "incorrect" displayed on screen for 1 second. As elaborated by the original authors, the experimenter will avoid evaluative behaviors during the pretest, to avoid causing the participant to feel as though they are being watched.

After the pretest, half of the participants will receive the "high-pressure" scenario, while the other half will receive the "low-pressure" scenario, both scripted on pp. 5–7 of the Supplementary Material. Those participants in the high-pressure scenario will make up our main study conditions, while those in the low-pressure scenario will make up our manipulation-check-comparison conditions. In the high-pressure scenario (but not the low-pressure scenario), they will be told that their improvement on the second half of the task will earn them and a (fictitious) partner additional money, and that their performance will be videotaped for teaching purposes. After delivering the high-pressure scenario, the experimenter will place a camera near the participant so as to record both the participant and their computer screen. The camera will not be started at this point in time.

In the *Control* condition, after the pretest, participants will be told to wait quietly for a few minutes while the experimenter retrieved some materials for later. In the *Expressive Writing* condition, participants will receive an envelope with writing instructions inside. The experimenter will tell them that they have a 10-minute writing session, and then leave the room. The envelope will contain instructions to write, as openly as possible, about their thoughts and feelings about the math problems. The instructions will clarify that nobody would ever be able to link up their responses with their id [full script on p. 8 of the Supplementary Material].

After the writing/control period is finished, in the high-pressure scenario, the experimenter will return to the room, start the camera, tell the participants that the camera is on, point to the red flashing light, and reinforce the high-pressure scenario manipulation, using the language on p. 9 of the Supplementary Material. Although the camera will be on, deception will occur, since the camera will not actually be recording the participant. All participants are then given 40 additional modular arithmetic problems (the post-test), similarly distributed between easy and difficult, as in the pretest, then the computer program will present all participants with the STAI. The program will also present those in the Writing condition with the Expressivity scale. All participants will be debriefed and paid for participation.

Analysis

The analysis will be performed exactly as in the supplemental materials (p. 10-13). Any participants who score below-chance on the pretest will be excluded (following the exclusion criteria on p. 2 of the Supplemental Material). No other exclusion rules were identified. We will include all other participants that complete at least a portion of the dependent variable.

We will only analyze performance on the 20 high-demand modular arithmetic problems, following the analyses of p. 10 of the supplemental materials. The critical test will measure differences in the percentage of the

high-demand modular arithmetic problems solved, in the high-pressure scenario condition, looking at the interaction between scores in the pretest vs post-test by the writing condition (expressive vs. control) using a 2 mixed within (test: pre vs. post)/between (writing: expressive vs. control)-subjects ANOVA. A follow-up between-subjects t-test will compare scores in the post-test between the expressive and control conditions, and two within-subjects t-tests will separately compare changes in performance from pretest to posttest separately in the expressive writing and control conditions. We will not look at performance in the low-pressure conditions.

As a manipulation check, we will compare STAI (anxiety) scores between all participants in the high-pressure scenario conditions and all participants in the low-pressure scenario conditions, using a between-subjects t-test (as in p. 13 of the Supplemental Materials).

The result 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 collection to test if the original result replicated (a two-sided p-value < 0.05 in the same direction as the original study).

As secondary analyses, we will look the correlation between Expressivity scores and improvement from pre- to post-test in the high-pressure expressive-writing condition.

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 with University of Virginia students between September 2016 and September 2017, whereas the data in the original study was carried with University of Chicago students, date unknown. As such, as in all replications, the sample, recruiting, and setting are different from the original study. There are no claims in the original article that suggest that these deviations are material for the tested effects. Nevertheless, following feedback from the original authors, we are restricting recruiting to the highest performing University of Virginia students, based on SAT or ACT scores.

Additionally, while the authors have provided scripts and guidance for inducing pressure in the high-pressure scenario, there will still be unavoidable differences in how the pressure-manipulation is delivered, which may create deviations in the amount of pressure felt by participants. We will attempt to minimize those differences before data collection by sending videos of the manipulation to the original authors for review, and after data collection by measuring whether the high-pressure scenario created more anxiety in participants than the low-pressure scenario. The computer program used to present participants with the modular arithmetic problems will also present the STAI (all participants) and the expressivity scale (writing condition participants only) at the conclusion of the post-test.

The original paper contains five studies: the replication is of study 1, following the project protocol to select the first study in the paper reporting treatment effects.

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

Participants. 57 students at the University of Virginia participated in the first data collection for course credit. All were recruited from the Department of Psychology's Participant

Pool, and all reported having received SAT scores > 1400 or ACT scores > 30. 69% of the participants were female, with an average age of 18.65 years ($SD = 1.03$). Five participants were excluded from analysis: one for a wrist injury (creating an impediment for their previously assigned writing condition), one for a participant that proceeded to the postblock without researcher confirmation during the 10 minute control condition break, and three for failing to score above chance on the practice modular arithmetic problems, leaving a final sample of 52 participants: 26 in the high-pressure conditions and 26 in the low-pressure conditions.

We found that the pressure manipulation was marginally successful in increasing anxiety (STAI, $alpha = 0.94$) among participants. Participants given the high-pressure instructions were more anxious ($M = 40.65$, $SD = 10.40$) than were those given the low-pressure instructions ($M = 35.85$, $SD = 10.19$); $t(50) = 1.6838$, $p = 0.10$, $d = 0.47$ [−0.10, 1.03] (as compared to a $d = 0.99$ in the original paper).

The focal hypothesis, however, was not supported by the data. Participants in the high-pressure conditions showed no difference in their math performance from pretest to posttest based on whether or not they had expressively written. $F(1, 24) = 0.1352$, $p = 0.72$. The direction of the effect was opposite to the equivalent test in the original study.

Prespecified follow-up tests indicated that those in the high-pressure condition who wrote expressively did not perform better on the posttest than those who did not: M(Expressive) = 17.85/20 correct, $SD = 1.77$; M(Control) = 17.92/20 correct, $SD = 1.71$; $t(24) = 0.1128$, $p = 0.91$, $d = 0.04$ [−0.77, 0.80]. The direction of the effect was opposite to the equivalent test in the original study.

Participants in the high-pressure control condition did not show a significant decrease in performance from pretest to posttest: M(Pretest) = 17.85/20 correct, $SD = 1.82$, M(Posttest) = 17.92/20 correct, $SD = 1.71$; $t(12) = -0.1273$, $p = 0.90$, $dz = -0.03$ [−0.58, 0.50]. The direction of the effect was opposite to the equivalent test in the original study.

Participants in the high-pressure expressive-writing condition did not show a significant change in performance from pretest to posttest: M(Pretest) = 18.08/20 correct, $SD = 1.26$, M(Posttest) = 17.85/20 correct, $SD = 1.77$; $t(12) = -0.3985$, $p = 0.70$, $dz = -0.11$ [−0.65, 0.44]. The direction of the effect was opposite to the equivalent test in the original study.

Secondary analyses looking at whether scores in the Expressivity scale ($alpha = 0.74$) correlated with improvement from pretest to posttest in the high-pressure expressive writing condition found no significant relationship: $r(11) = 0.10$ [−0.62, 0.47], $p = 0.73$.

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

Participants. 138 students at the University of Virginia participated in the pooled collection for course credit. All were recruited from the Department of Psychology's Participant Pool, and all reported having received SAT scores > 1400 or ACT scores > 30. 68% of the participants were female, with an average age of 19.17 years ($SD = 3.32$). In addition to the five participants excluded in the first data collection, an additional two participants were excluded in the second data collection: one due to the participant becoming aware of their assigned condition prior to being run (entering the room prematurely while the experimenter was still completing exper-

imental setup), and one for failing to score above chance on the practice modular arithmetic problems. Their exclusion leaves a final sample of 131 participants: 79 in the high-pressure conditions and 52 in the low-pressure conditions.

We found that the pressure manipulation was successful in increasing anxiety (STAI, $alpha = 0.94$). Participants given the high-pressure instructions were more anxious ($M = 42.22$, $SD = 11.81$) than were those given the low-pressure instructions ($M = 36.87$, $SD = 10.05$); $t(129) = 2.7780$, $p = 0.006$, $d = 0.50$ [0.14, 0.85].

The focal hypothesis, however, was not supported by the data. Participants in the high-pressure conditions showed no difference in their math performance from pretest to posttest based on whether or not they had expressively written. $F(1, 77) = 0.7352$, $p = 0.39$. The direction of the effect was opposite to the equivalent test in the original study.

Prespecified follow-up tests indicated that those in the high-pressure condition who wrote expressively did not perform better on the posttest than those who did not: M(Expressive) = 17.71/20 correct, $SD = 1.87$; M(Control) = 17.58/20 correct, $SD = 2.18$; $t(77) = 0.2808$, $p = 0.78$, $d = 0.06$ [−0.39, 0.52]. The direction of the effect was the same as the equivalent test in the original study.

Participants in the high-pressure control condition did not show a significant decrease in performance from pretest to posttest: M(Pretest) = 17.87/20 correct, $SD = 2.04$, M(Posttest) = 17.58/20 correct, $SD = 2.18$; $t(44) = 1.1221$, $p = 0.27$, $dz = 0.17$ [−0.13, 0.46]. The direction of the effect was the same as the equivalent test in the original study.

Participants in the high-pressure expressive-writing condition did not show a significant change in performance from pretest to post-test: M(Pretest) = 18.38/20 correct, $SD = 1.21$, M(Posttest) = 17.71/20 correct, $SD = 1.87$; $t(33) = -1.7206$, $p = 0.10$, $dz = -0.30$ [−0.64, 0.05]. The direction of the effect was opposite to the equivalent test in the original study.

Secondary analyses looking at whether scores in the Expressivity scale ($alpha = 0.59$) correlated with improvement from pre-test to post-test in the high-pressure expressive writing condition found no significant relationship: $r(32) = 0.04$ [−0.30, 0.37], $p = 0.82$.

Unplanned Protocol Deviations

We inadvertently oversampled high-pressure conditions and undersampled low-pressure conditions in the second data collection – we targeted 33/cell for each of the four conditions, and instead ended up with 25 in the low-pressure control, 27 in the low-pressure expressive writing, 45 in the high-pressure control, and 34 in the high-pressure expressive writing conditions at the conclusion of the Spring 2017 academic semester.

All conditions for both the first and second rounds of collection were randomly assigned using a random number generator (found at Random.org), to generate numbers 1–4, with each digit assigned to one of the 4 experimental conditions. This created a random imbalance between the two high pressure conditions over the course of the pooled first and second round collections. Additionally, prioritization was placed on reaching the target sample for the main high pressure conditions used in the focal analysis during the Spring 2017 collection period, which led to an imbalance in the manipulation-check low pressure conditions. The sample that we had already collected (79 in the high-pressure conditions, and 52 in the low-pressure conditions) gave us better than 99.9% power to detect the original manipulation-check effect size of $d = 0.99$, and 79% power to detect 50% of that effect size. Also, because we oversampled the high-

pressure conditions, we had more power to detect the focal tests than prespecified.

Discussion

Despite high power and careful adherence to the original study design, with feedback from the original authors, we did not observe an effect consistent with the original study.

Our sample seems roughly equivalent to that used in the original, at least in terms of math ability: participants in the replication performed similarly to the original study on solving math problems before the intervention, M(replication) = 90% correct; M(original) = 82% correct, and participants in the replication were also similar to the original study on solving math problems in the low-pressure conditions after the interventions, M(replication) = 90% correct, M(original) = 84% correct.

Our successful manipulation check suggested that we induced the feelings of pressure necessary to elicit the effect. Participants in the replication felt significantly more anxiety in the high-pressure conditions than in the low-pressure ones: $d = 0.50$ [0.14, 0.85], albeit less so than in the original study ($d = 0.99$). Despite these consistencies in overall performance and effectiveness of the manipulation, the replication showed no difference in post-test performance between those in the high-pressure expressive writing and control conditions ($d = 0.06$ [−0.39, 0.52]; original study $d = 2.48$). The original effect size did not fall within the confidence interval of the replication. Additionally, we found the direction of our effect was opposite that of the equivalent tests in the original study for both the first round of collection as well as the first and second rounds of collection pooled.

This failure to replicate suggests caution about the reliability of the original result, but does not definitively suggest that the original result was a false positive. There could be as yet unidentified differences between the original and replication methodology that are critical for observing this effect.

References

Ramirez, G. / Beilock, S. L. (2011): *Writing About Testing Worries Boosts Exam Performance in the Classroom*, Science, 331, pp. 211–213.