Replication of Bear et al. 2017
"Mistiming of thought and perception predicts delusionality"
PNAS 114(40), 10791-10796.
https://www.pnas.org/content/114/40/10791

The original paper only includes one study. In this experiment, participants perform a prediction task where they see five empty squares in random locations on a screen and predict which squares will light up red. After that, participants indicate whether they have made a correct prediction, an incorrect prediction, or no prediction. Participants also complete a control task measuring participants' general ability to discriminate two external visual events in time, and after that, they answer questions on the 21-item Peters et al. Delusions Inventory (PDI) that measures a general proneness to a broad area of delusion-like ideation. The authors manipulate the amount of time participants have to make their prediction before a square is randomly chosen and turns red. Postdiction in the prediction task is assessed by exploring participants' rate of claiming to have made an accurate prediction among trials in which they think they completed their prediction before the square turned red as a function of the time they were given to make this prediction, with more biased predictions (more postdiction) being more likely in shorter trials. Participants who score higher on the PDI have an exaggerated bias in shorter trials.

Hypothesis to replicate and bet on: People more prone to delusion-like ideation have an exaggerated bias in accurate prediction in shorter trials. To evaluate this hypothesis, the authors use a multilevel logistic model, with a focus on the interaction effect of delay with PDI ($b = -0.013$, $z = -2.94$, $p = 0.003$); p.10793.

Criteria for replication: The criteria for replication are an effect in the same direction as the original study and a p-value < 0.05 in a two-sided z-test.

Power analysis: The original study had 994 participants (19 participants were excluded since they admitted to "cheating" on the prediction task). The standardized effect size (Cohen's d) was $d = 0.187$. To have 90% power to detect 67% of the original effect size, a sample size of $n = 2719$ is required.

Sample: The original paper excludes participants who answered at least one of three comprehension questions incorrectly twice. Trials when participants answered "didn't have time" to do the prediction were excluded from the study. Participants were also excluded if they admitted to "cheating" on the prediction task in the debriefing survey. We will use the same exclusion criteria. We will make sure that participants can only participate once from the same account in this specific study, and we will only recruit participants with a HIT approval rate of 95% or above. We will also check all IP addresses via https://www.ipqualityscore.com/; and we will remove any participants where one or more of the following is true: fraud score >= 85; TOR = True; VPN = True; Bot = True; abuse velocity = high. The replication sample size is the sample size after any exclusions of participants.

Materials: We will use the same material as in the original study, kindly provided by the original authors. In particular, the experiment will be conducted using the original *Qualtrics* survey.

Procedure: We will closely follow the procedure of the original experiment. The following summary of the experimental procedure is therefore largely based on the description of the experiment in the article (p. 10792) and the Supplementary Information (pp. 1–3).

Participants will first be shown a Captcha, and will thereafter provide informed consent. After this we will include an attention check that participants will need to pass to continue to the study. This attention check is in addition to any other potential attention check(s) used in the original study. Participants will then be asked to complete a prediction task with 140 trials in which they will presented with a fixation cross [a 30-pixel (px) +] for 500 ms, followed by five empty squares that will appear in random locations on the screen. Participants will be asked to predict which of these squares will light up red; they will be tasked to "pick (in your head) a single square that you think will turn red" before one of the squares is randomly selected to turn red. The position of these five squares will be randomly chosen on each trial from a set of possible locations of a 5 × 5 grid, centered on the position of the previously shown fixation cross. To force participants to attend outside of the place of fixation, a square will never be presented directly where the fixation cross has been presented (so there are only 24 possible locations where the squares can appear). Each square will have 50 × 50 px dimensions, and the 5 × 5 grid of possible locations will span a region that is 330 × 330 px (leaving 20 px spacing between each possible square location). Participants will then be asked to indicate whether they have made a correct prediction, an incorrect prediction, or no prediction (if they do not have time to finish before one of the squares lights up).

We will manipulate the amount of time participants are given to make their prediction before one of the squares is randomly selected to turn red. The delay between the initial presentation of the empty squares and the moment at which one of these squares is selected to turn red will be experimentally manipulated. There will be seven possible delays: approximately 100, 150, 250, 400, 600, 2,000, and 4,000 ms with each type of delay on 20 trials, presented randomly across the 140-trial sequence.

Before the task begins, participants will be given detailed instructions on how to perform the task. To avoid encouraging participants to report that they make a prediction before a square turns red even on the extremely fast trials for which this is difficult, the instructions will emphasize that, although participants should try to make their predictions as quickly as possible, "it's completely understandable and expected that you won't always have a chance to complete your guess in time, even if this is the case on most or all trials." Participants will also be instructed to "try to make your guess right when the empty squares appear and no earlier."

Participants will need to answer three comprehension questions correctly to proceed. Those who make a mistake will be given one more chance to read the instructions and answer these questions correctly; and if they fail to answer one of the questions correctly a second time, they will be excluded from the study.

Participants who pass the comprehension check will complete seven practice trials before the main task, consisting of a single trial of each of the seven possible delays, presented in a random order.

Participants will also complete a control task with 140 trials (temporal discrimination task, in a counterbalanced order with the prediction task) that measures participants' general ability to

discriminate two external (visual) events in time. The task is similar in appearance to the prediction task, but instead of making a prediction, participants will indicate whether a square flashing red is preceded or followed by the screen blinking (going blank for a brief moment). We will randomly select the order of these events and manipulate the delay between them. Also for this task, there are seven practice trials. In the original study, the authors test whether participants endorsing more delusion-like ideation are worse at temporal discrimination – the opposite is the case in the original study. We will also test whether this is the case. The conclusion about whether the study replicates or not will only be based on the main replication test (i.e., the results of the test from the temporal discrimination task will not affect this conclusion). The test from the temporal discrimination task is mainly relevant for understanding why the study failed to replicate if it should fail to replicate.

After these two tasks, participants will be administered the 21-item Peters et al. Delusions Inventory (PDI), which measures a general proneness to a broad array of delusion-like ideations (e.g., paranoia, magical thinking, reference). PDI scores will be log-transformed as in the original study.

At the end of the study, participants will be debriefed and those that admit to "cheating" on the prediction task will be excluded. Trials in which participants answered that they "didn't have time" to make the prediction will also be excluded.

Analysis: The analysis will be performed as in the original paper, with the use of a multilevel logistic model that includes participant-specific random intercepts, with a focus on the interaction effect of delay with PDI. Given the replication team's concerns about the model's underlying assumptions, we will conduct two additional secondary tests. The first test will rely on a model that also includes participant-specific random slopes for delay, given that the primary model is not maximally specified (such a model was also reported in the original paper as a secondary analysis). The second test will rely on a linear (rather than logistic) multilevel model that includes participant-specific random intercept and slope for delay, given that the interpretation of interaction coefficients in logistic models is not straightforward.

Subject payments: We are standardizing payments across all replications so that studies have a certain show-up fee depending on the expected length of the study, with an hourly wage from the show-up fee of $8 and a minimum payment of $1 (for studies with incentive payment we use the same incentive payment as in the original study; and this payment is paid in addition to the show-up fee). If we have problems recruiting, we will increase the show-up fee.