Replication of Handley et al., 2015
"Quality of evidence revealing subtle gender biases in science is in the eye of the beholder"
PNAS 112 (43), 13201-13206.
https://www.pnas.org/content/112/43/13201

The original paper includes two MTurk studies but only experiment 3 fits our replication criteria. In this between-subject experiment, participants are randomly assigned either a treatment where they read and evaluate an article abstract reporting a gender bias against women in scientific conference submissions, or to a treatment where they read and evaluate an altered abstract reporting no gender bias. Men evaluate the abstract reporting that gender-bias exists less favorably than women do and the abstract reporting that no gender-bias exists more favorably than women.

Hypothesis to replicate and bet on: Men evaluate an abstract reporting that gender-bias exists less favorably than women and an abstract reporting that no gender-bias exists more favorably than women. To evaluate this hypothesis, the authors perform an *F*-test on the gender x abstract type interaction ($F(1,299) = 4.00$, $p = 0.046$, $\eta^2_{partial} = 0.013$); p. 13203.

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

Power analysis: The original study had 303 participants. The standardized effect size (Cohen's *d*) was $d = 0.460$. To have 90% power to detect 67% of the original effect size, a sample size of $n = 1791$ is required.

Sample: Only participants from the US over 18 years of age were allowed to participate in the original study. While 321 individuals participated in the experiment, 12 participants were excluded from data analysis because they failed one or more attention-check items, 2 because they reported being under 18 years of age or did not specify an age, 1 because they did not specify their gender, and 7 because they reported they had read the abstract before (the authors report that some participants met multiple exclusion criteria). 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.

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 Supplementary Information (pp. 1–2).

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 start by reading the following paragraph:

In the scientific world, peers often judge the quality of research and decide whether or not to publish it, fund it, or discard it. We are conducting an academic survey about people's opinions about different types of research that was published in the last few years. You will be asked to read one randomly selected abstract and asked to provide your opinion. This is akin to reading an abstract for consideration in a conference symposium or in consideration of whether to send out a paper for further review. There is no right or wrong answer and we realize you don't have all the information or background. We are especially interested in what types of research faculty think is compelling versus needs more empirical support. Just like in the scientific world, many judgments are made on whether something is quality science or not after just reading a short abstract summary. So, to create that experience for you, we ask that you provide your overall reaction as best you can even with the limited information. You will also be asked to provide demographic information about yourself for statistical information only. Please read the following abstract from a 2012 published research study and then provide your opinion with the items below.

Participants will then be randomly assigned to read either the original version of the Knobloch-Westerwick et al. (2013), which reported a gender bias (e.g., "Publications from male authors were associated with greater scientific quality, in particular if the topic was male-typed"), or a version slightly altered to report no gender bias (e.g., "Publications from male and female authors were associated with comparable scientific quality, even if the topic was male-typed").

Participants will be asked to evaluate the abstract by answering four questions, on a scale from 1 (*not at all*) to 6 (*very much*):*"To what extent do you agree with the interpretation of the research results?", "To what extent are the findings of this research important?", "To what extent was the abstract well written?", and "Overall, my evaluation of this abstract is favorable."* The outcome measure of interest is the average of these four responses.

Finally, participants will be asked to complete demographic information, and will be debriefed regarding the purpose of the experiment.

Analysis: The analysis will be performed as in the original article. In particular, we will perform a 2 (gender: male or female) x 2 (abstract type: original or modified) ANOVA. The replication focuses on the *F*-test on the gender x abstract type interaction.

Subject payment: 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.