I have gladly accepted the authors' invitation to share my thoughts on the current manuscript. The supporting analyses are exclusively based on the data from this article. Given the limits of time and space, no further inquiries using the original data set were conducted.

First of all, let me laude the replicators' effort in this extensive enterprise. I admit having been very surprised that the original finding was not obtained. Originally, I had suggested the "pen study" to be replicated not only because we (SMS) had found the predicted difference but also because numerous operational and conceptual replications from our and our colleagues' labs have confirmed the original results. During the past five years, at least 20 studies have been published demonstrating the predicted effect of the pen procedure on evaluative judgments (For a selection of relevant publications since the year 2000, see https://www.dropbox.com/s/5ttmh4swuhwgs17/Literature.xlsx?dl=0).

Because the authors have neither formulated specific hypotheses nor created conditions focusing on moderating factors, speculating about the reasons of this failure is necessarily ex post. However, there are several aspects of the current replication endeavor that deserve attention.

First, the authors have pointed out that the original study is "commonly discussed in introductory psychology courses and textbooks". Thus, a majority of psychology students was assumed to be familiar with the pen study and its findings. Given this state of affairs, it is difficult to understand why participants were overwhelmingly recruited from the psychology subject pools. The prevalent knowledge about the rationale of the pen study may be reflected in the remarkably high overall exclusion rate of 24 percent. Given that there was no funneled debriefing but only a brief open question about the purpose of the study, to be answered in writing, the actual knowledge prevalence may even be underestimated.

That participants' knowledge of the effect may have influenced the results is reflected in the fact that those 14 (out of 17) studies that used psychology pools gained an effect size of $d = -0.03$ with a large variance ($SD = 0.14$), while the three studies using other pools (Holmes, Lynott, and Wagenmakers) gained an effect size of $d = 0.16$ with a small variance ($SD = 0.06$). Tested across the means of these studies, this difference is significant, $t(15) = 2.35$, $p = .033$; and the effect for the non-psychology studies significantly deviates from zero, $t(2) = 5.09$, $p = .037$, in the direction of the original result.

Second, and despite the obtained ratings of funniness, it must be asked if Gary Larson’s The Far Side cartoons that were iconic for the zeitgeist of the 1980s instantiated similar psychological conditions 30 years later. It is indicative that one of the four exclusion criteria was participants’ failure to understand the cartoons.

Third, it should be noted that to record their way of holding the pen, the RRR labs deviated from the original study by directing a camera on the participants. Based on results from research on objective self-awareness, a camera induces a subjective self-focus which may interfere with internal experiences and suppress emotional responses.

Finally, there seems to exist a statistical anomaly. In a meta-analysis, when plotting the effect sizes on the x-axis against the sample sizes on the y-axes across studies, one should usually find no systematic correlation between these two parameters. As pointed out by...
Shanks and his colleagues (e.g., Shanks et al., 2015), a set of unbiased studies would produce a pyramid in the resulting funnel plot such that relatively high-powered studies show a narrower variance around the effect than relatively low-powered studies. In contrast, an asymmetry in this plot is seen to indicate a publication bias and/or p-hacking (Shanks et al., 2015).

For the present 17 studies, the funnel plot looks like this:

![Funnel Plot](image)

Obviously, there is no pyramidal shape of the funnel. Although all of the present studies were appropriately powered, there were several relatively low-powered unsuccessful studies (at the left bottom) but only a few relatively low-powered successful studies (at the right bottom). This is reflected in a relationship between effect size and sample size, \( r(17) = .45, p = .069 \), such that the size of the sample is positively correlated with the size of the effect. Note that this pattern is the opposite of what is usually interpreted as a publication bias or p-hacking in favor of an effect (resulting in a negative correlation between effect size and sample size). Without insinuating the possibility of a reverse p-hacking, the current anomaly needs to be further explored.

In summary, while a first look at the current data seems to suggest that the SMS facial-feedback study has been convincingly “nonreplicated”, a closer inspection of the replication studies reveals several methodological and statistical issues that need to be considered before drawing further conclusions on the validity of the method, of the model, or of the underlying mechanism.

Reference