When Both the Original Study and Its Failed Replication Are Correct: Feeling Observed Eliminates the Facial-Feedback Effect

Tom Noah, Yaacov Schul, and Ruth Mayo
Hebrew University of Jerusalem

This article suggests a theoretically driven explanation for a replication failure of one of the basic findings in psychology: the facial-feedback effect. According to the facial-feedback hypothesis, the facial activity associated with particular emotional expressions can influence people's affective experiences. Recently, a replication attempt of this effect in 17 laboratories around the world failed to find any support for the effect. We hypothesize that the reason for the failure of replication is that the replication protocol deviated from that of the original experiment in a critical factor. In all of the replication studies, participants were alerted that they would be monitored by a video camera, whereas the participants in the original study were not monitored, observed, or recorded. Previous findings indicate that feeling monitored or observed reduces reliance on internal cues in making judgments. Therefore, we hypothesize that recording the participants in the replication experiments reduced their reliance on the facial-feedback. To test the hypothesis, we replicated the facial-feedback experiment in 2 conditions: one with a video-camera and one without it. The results revealed a significant facial-feedback effect in the absence of a camera, which was eliminated in the camera's presence. These findings suggest that minute differences in the experimental protocol might lead to theoretically meaningful changes in the outcomes. In our view, the theoretical and methodological approach advocated by our study changes failed replications from being the "end of the road" regarding entire fields of study into a new road for growth regarding our understanding of human nature.

Keywords: camera presence, context effects, facial-feedback, feeling observed, replications in social psychology

Supplemental materials: http://dx.doi.org/10.1037/pspa0000121.supp

A central question in social psychology concerns how the presence of others affects human judgment and behavior. Numerous studies reveal that people behave differently when they feel observed and when they experience privacy (Forgas, Brennan, Howe, Kane, & Sweet, 1980; Froming, Walker, & Lopyan, 1982; Triplett, 1898; van Bommel, van Prooijen, Elffers, & Van Lange, 2012; Zajonc, 1965). In part, these behavioral changes reflect attempts of protagonists to tailor their reactions to the presence of the observers to maximize their outcomes. Protagonists may strategically try to increase their gains by, for example, using impression management techniques (Schlenker, 1980) or adapting their argumentation on the basis of their analysis of the others (e.g., Eagly, Wood, & Chaiken, 1978). Importantly, however, protagonists' behaviors while being observed may reflect spontaneous reactions to the presence of others that may not involve conscious attempts to please the observers. To illustrate, recently we found that when people feel observed, they rely less on their metacognitive feelings in judgments and decisions (Noah, Schul, & Mayo, in press). Such an effect can occur because when feeling observed, people adopt an external perspective of themselves (Hass, 1984; Wiekens, 2009). Accordingly, they tend to neglect internal information (Scheier & Carver, 1980; Wicklund & Duval, 1971) and base their judgment mainly on external information that is visible to potential observers (Steinmetz, Xu, Fishbach, & Zhang, 2016). In the current research we suggest that the diminished reliance on internal cues when one feels observed can inform the recent discussion about the facial-feedback effect (Strack, Martin, & Stepper, 1988) and its failed replication attempts (Wagenmakers et al., 2016).

According to the facial-feedback hypothesis, the facial activity associated with particular emotional expressions can influence people's affective experiences. In a classic study of the effect, Strack and colleagues (1988) asked participants to view amusing cartoons while holding a pen either between their teeth or between their lips. It was assumed that the former facial posture contracts

Tom Noah, Department of Psychology and the Federmann Center for the Study of Rationality, Hebrew University of Jerusalem; Yaacov Schul and Ruth Mayo, Department of Psychology, Hebrew University of Jerusalem.

We thank Linor Kagan, Neta Galperin, Hagai Gumpert, Inbar Amgar, Hadas Rozett, Yonatan Milson Dagan, Bat-El Terehovsky, Shir Gabay, Tomer Kupershmidt, and Elena Balchugov for their help in running the experiment. We also thank Maya Bar-Hillel, Moran Sela, and Henry Zukier for their useful feedback.

Correspondence concerning this article should be addressed to Tom Noah, Department of Psychology and the Federmann Center for the Study of Rationality, Hebrew University of Jerusalem, Mount Scopus, Jerusalem 9190501, Israel. E-mail: tom.noah@mail.huji.ac.il

the zygomaticus major muscle that is used in smiling, and the latter contracts the orbicularis oris muscle, which inhibits muscle activity associated with smiling. Strack et al. reported that the participants who held the pen between their teeth rated the cartoons as funnier than did participants who held the pen with their lips. Those findings were interpreted as indicating that the facial activity associated with an emotional expression can influence people's affective experiences even when they are not consciously aware of their facial expressions. In a recent review, Laird and Lacasse (2014) concluded: "the basic notion that emotional feelings are consequences of expressions and autonomic responses has been supported over and over" (p. 31).

Notwithstanding, Wagenmakers et al. (2016) recently published a registered replication report describing the results of 17 direct replications of the original Study 1 of Strack and colleagues (1988). Contrary to the findings in the original study, none of the 17 replication experiments revealed support for the facial-feedback effect using the pen-in-the-mouth paradigm, leading the authors to question the validity of the original findings.[1] Strack (2016) noted that the protocol of the replication experiments deviated from the original paradigm in several ways, one of which was placing a video camera in front of each participant, and informing the participant that the task performance will be video-recorded to verify that the pen is held correctly throughout the experimental tasks.

Video-recording experimental sessions is a common practice in current psychological research. It is often considered as a mean to improve the quality of data collection and analysis, because it provides documentation of otherwise not accessible information regarding participants' behavior during the session.[2] However, the theoretical analysis briefly reviewed above suggests that the presence of the camera might be critical for the facial-feedback effect. Accordingly, we propose that the two sets of findings, namely the presence of the facial-feedback effect under the conditions of the original experiment (Strack et al., 1988) and its absence under the conditions of the replication experiments (Wagenmakers et al., 2016), are not contradictory. Rather, they are consistent with the influence of feeling observed on people's reliance on their inner cues. To test our hypothesis, we investigated the facial-feedback effect using the paradigm of the replication project in two conditions: without a camera (as in the original study), and with a camera in front of the participants during the experiment (as in the replication project). We hypothesized that participants would rely on the internal cues associated with the facial feedback when they do not feel observed (i.e., in the no-camera condition), but not when they do feel observed (i.e., in the camera-present condition).

Notwithstanding the specific research question regarding the impact of feeling observed on the facial-feedback effect, we believe that the current experimental design offers a method for handling the doubts raised by failed replications. This adds to recent studies exploring the influence of context on replicability. To name two prominent examples, Van Bavel, Mende-Siedlecki, Brady, and Reinero (2016) demonstrated that sensitivity of effects to contextual factors (time, culture, location, and population) was negatively related to success in replication attempts, and Luttrell, Petty, and Xu (2017) compared between running the same study under optimal and nonoptimal conditions (which were used, respectively, in the original study and its failed replication attempt). The two conditions in Luttrell et al.'s study (2017) differed in several key features, such as stimulus length, its relevance for the participants, and using the full versus short version of a scale. Luttrell et al. found the original effect under the optimal condition, but not under the nonoptimal condition. This finding was later replicated in an additional study by the authors of the replication attempt (Ebersole et al., 2017). We propose that it is vital to examine cases of replication failure within a theoretical and methodological framework that combines both the original study and its replication within the same experimental array, altering only one theoretically meaningful factor. Such investigation allows a test of the theoretically derived factor that is predicted to lead to an effect in one condition but not in the other. We hope that isolating and testing moderators' influence on an effect can help not only in reconciling contradicting results, but also in identifying phenomena that might be interesting and important, both theoretically and methodologically.

Method

Experimental Design

The current study explores whether the presence of a camera during the experiment affects the facial-feedback effect. The experiment had two between-participants factors (Camera Presence [yes/no] × Facial Activity [lips/teeth]). The DV was the average amusement ratings of the four cartoons.

The sample size was determined according to the procedure recommended by Cohen (1988), with the aid of G-Power software (Version 3.1.9.2; Faul, Erdfelder, Lang, & Buchner, 2007). Specifically, it was based on an estimate of the effect size of Experiment 1 by Strack et al. (1988); ($SD = 1.69$, $d = .49$), and was set for a significance level of $\alpha = .05$ and a power of 80%. We note that the estimate of the within-condition σ derived from the Strack et al.'s study, is similar to that derived from the 17 replications in Wagenmakers et al.'s study (average $SD = 1.51$). We rounded the result to the closest multiple of 10, resulting in 50 participants per condition, which is the minimum sample size in the replication project (Wagenmakers et al., 2016). We also note that in choosing the sample size we had to decide between two goals: one was to test whether the simple effect of facial feedback replicates under the same conditions as in the replication project except for the camera presence (i.e., using a similar sample size with similar statistical power); the other goal was to have sufficient statistical power for detection of an interaction effect (involving the camera presence/absence and the facial-activity manipulation). The difference between the lips-teeth contrasts in the original study and in the replication (the interaction effect) is about .80. Based on G-Power, detecting such an interaction requires more than 120 participants per condition. We opted to align the number of participants per condition with the replication study, namely, based on

[1] Although in this article we refer to the validity of the facial-feedback effect, it is important to note that, as also acknowledged by Wagenmakers et al. (2016), the registered replication does not challenge the general effect of facial feedback, but rather the specific pen-in-the-mouth paradigm by Strack et al. (1988), which was designed as a specific unobtrusive method for testing the facial-feedback effect.

[2] Indeed, according to personal communication with E.J. Wagenmakers (January 2018), the use of a camera during the replication experiments was not initiated by the authors who conducted the research but rather by a facial-feedback expert who reviewed the protocol in advance.

the power to detect the simple effects. Although the interaction is useful for showing the role of the presence of a camera in this paradigm, it is not critical for our argument. That is, our argument is not about the specific effect of the presence of a camera, but rather, about the consequences of its absence—the existence of the facial feedback effect.

We uploaded the method and analysis plan in advance to the Open Science framework (https://osf.io/xt39q). This is the only experiment we ran in order to test the effect of feeling observed on the reliance on facial feedback, and we report here the complete results according to the preregistration.

Participants

Two hundred participants (54% females, mean age 24, $SD = 2.32$) completed the experiment in exchange for payment of approximately US$5 (in Israeli currency) or course credit. The experiment was advertised as suitable only for native Hebrew speakers who are not psychology students. This was done to minimize the number of participants who are familiar with the research on the facial-feedback hypothesis.

Materials and Procedure

The materials and procedure of the study were based on the procedure of the replication project (Wagenmakers et al., 2016). The instructions to experimenters, information form, consent form, instructions to participants, and task booklet were translated into Hebrew from the replication materials that were uploaded to the Open Science framework (for a detailed description, see supplementary). Two notable changes were made: (a) We replaced the cartoons that were used in the replication study with four cartoons in Hebrew that were chosen based on the findings of a pretest on an Israeli sample[3]; (b) In the no-camera conditions there was no camera in the experimental room, and the experimenter did not tell the participants that they would be videotaped while completing the tasks in the booklet.

Each participant was randomly assigned to one of the four experimental conditions and underwent the experiment alone in an isolated room, behind a closed door. As can be seen in Figure 1, each room was furnished with a desk and computer, and the participant was supplied with the information and consent forms, a Stabilo pen 68 mm, an alcohol swab, and a box of paper tissues. In the camera condition, a video camera was connected to the computer and located right in front of the participant's chair. The setup in the no-camera condition was identical, except that the camera was absent.

Upon arrival to the lab, participants read the information brochure and signed the informed-consent form. Then they read the instructions from the computer screen. The instructions explained and demonstrated the right and wrong ways of holding the pen, according to the experimental condition (using the lips or the teeth). After verifying that the instructions were clear, the experimenter gave the participants the task booklet.

The booklet began with a practice task, which required drawing a straight line between two points while holding the pen in the requested position. The participant completed the practice task in the presence of the experimenter, who made sure that the participant was holding the pen correctly. In the camera condition, duplicating the procedure of the replication project (Wagenmakers et al., 2016), the experimenter explained to the participant that during the experiment a video recording is made to verify that the pen is held correctly throughout the task. Then the experimenter pressed the button that turned on the camera and made sure that it was directed toward the participant's face.[4] In both conditions, the experimenter left the room and closed the door behind him/her, so the participants completed the booklet alone.

The first nonpractice task in the booklet involved connecting 10 digits printed on the page by a line in a numerical order. A 10-point scale was printed on the bottom of the page, and participants were instructed to use this scale to indicate how difficult it was for them to perform the digit-connection task. This was done while holding the pen in their lips or teeth. The next page presented eight consonants and nine vowels. The participants' task was to underline only the vowels. After doing so, participants were again asked to rate the difficulty of the task on a 10-point scale printed on the bottom of the page. The third task was actually the task of interest, in which participants were told that they would see several cartoons of the kind typically found on the Internet and that, as usual, some would seem funnier than others. Participants were asked to rate the feeling that each cartoon induced on a 10-point scale ranging from *I felt not at all amused* (0) to *I felt very much amused* (9). Participants read and rated each cartoon with the pen held in the original position (i.e., lips or teeth).

After completing these tasks, the participants were asked to remove the pen from their mouth and answer 3 questions: (a) "How successful were you in holding the pen in the correct position during the entire experimental session?" (0–9 scale); (b) "Did you understand the cartoons?" (yes/no for each cartoon); and (c) "What do you think the purpose of this experiment is?" (open-ended). Participants were informed that those who correctly guess the aim of the experiment would enter a lottery with a prize of 100 NIS (about US$25). Then the participants provided their age, gender, status as a student (yes/no), and occupation or field of study. After participants completed the questionnaire, the experimenter asked whether they succeeded in holding the pen as requested, what facial expression they held during the experiment, whether they were familiar with the facial-feedback theory and effect, and whether they had thought about it during the experiment. Finally, the experimenter paid the participants and thanked them for their cooperation.

Exclusion of Participants

Based on the criteria specified in the preregistered plan, we excluded 26 participants from the analysis for the following four reasons: (i) One participant's average cartoon rating deviated from the average rating of all participants by more than 2.5 standard deviations; (ii) three participants correctly guessed the purpose of the study; (iii) two participants were not native Hebrew speakers; and (iv) 20 participants reported in the debriefing that they did not hold the pen as instructed during the tasks. These 20 participants were distributed as follows: three in the camera-teeth condition, eight in the no-camera-teeth condition, five in the camera-lips condition, and four in the no-camera-lips condition. In addition, we excluded eight participants

[3] The four cartoons were selected from various sources on the internet and had been pre-rated as being moderately funny.

[4] This procedure was designed to induce the feeling of being observed. However, in fact we did not record the sessions or observe them in any way. Therefore, unlike the replication project (Wagenmakers et al., 2016), participants were not excluded on the basis of actual compliance with the lips/teeth manipulation, in the camera and the no-camera conditions.

Figure 1. The experimental room in the camera condition (left) and in the no-camera condition (right). See the online article for the color version of this figure.

for the following three reasons that we did not anticipate in advance in the preregistered plan: (v) Four participants suspected the cover story of video recording (two in the no-camera condition who asked whether they were being recorded, and two in the camera condition who mentioned to the experimenter that the camera does not work or is not directed toward them); (vi) two participants did not agree to use the pen as instructed during the experiment proper (one insisted on using his own pen, and one wrapped the supplied pen with paper tissue); and (vii) two participants could not be included in the analysis because the experimenter neglected to record their experimental condition. Consequently, the effective sample size was 40–43 participants in each condition.[5] The Appendix contains a series of analyses that examine the sensitivity of our findings to the different criteria. Briefly, the main findings are robust to the exclusion criteria we use. In fact, the pattern of findings holds even when we include all of the participants in the analysis.

Results

Confirmatory Analysis

Our preregistered plan was to conduct a two-way between-participants analysis of variance (ANOVA; Camera presence [yes/no] × Facial activity [lips/teeth]) on the average amusement rating of the cartoons that were reported as understood. We hypothesized that based on the facial feedback effect (Strack et al., 1988) in the absence of a camera the cartoons would be rated as funnier in the teeth condition than in the lips condition. Moreover, we hypothesized that this effect will be smaller (or nonexistent) when a camera is present. In addition to the ANOVA, we computed the Bayes factors (BF)[6] using the Dienes calculator (Dienes, 2014). Figure 2 presents the mean ratings in the different conditions.

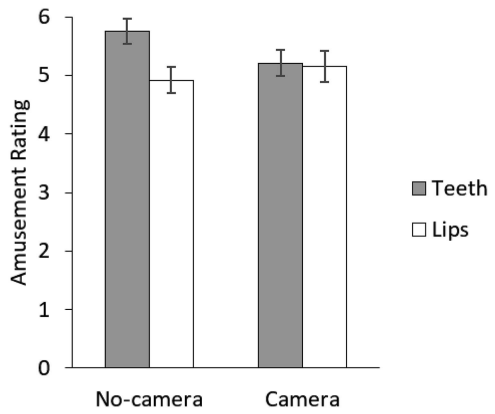

Figure 2. Means and *SE* of participants' amusement ratings (calculated only for cartoons that were reported as understood, $N = 166$).

According to the main hypothesis of this research, both analyses indicated that in the no-camera condition, as in the study by Strack et al. (1988), the participants in the teeth condition rated the cartoons as significantly more amusing than did participants in the lips condition ($M_{teeth} = 5.75$, $SD = 1.35$, $M_{lips} = 4.92$, $SD = 1.46$), $t(162) = 2.48$, $p = .01$, $\eta_p^2 = .038$, 95% confidence limits: [.171–1.500], BF = 8.66. In the camera-present condition, similar to the findings of Wagenmakers et al. (2016), this difference was much smaller and nonsignificant ($M_{teeth} = 5.21$, $SD = 1.46$, $M_{lips} = 5.15$, $SD = 1.74$), $t(162) = .19$, $p = .85$, $\eta_p^2 < .001$, 95% confidence limits: [−.586–.712], BF = 0.364. Are the two simple effects different from each other? Although the test of the 2 × 2 interaction was greatly underpowered, the preregistered analysis concerning the interaction between the facial expression and camera presence was marginally significant in the expected direction, $t(162) = 1.64$, $p = .051$, one-tailed, $\eta_p^2 = .016$, BF = 2.163. As predicted, the analysis did not reveal any main effect of camera presence or of facial feedback.

Exploratory Analysis

The above analysis is based only on the ratings of the cartoons that were reported as understood. That is, if a participant indicated that s/he did not understand one of the cartoons, that cartoon was not included in the average amusement rating of that participant. In

[5] In the pre-registration we declared that we plan to replace excluded participants by new ones. However, because we finished collecting the data from 200 participants on the last day of the academic year, we stopped collecting the data at this point, without replacing the excluded participants.

[6] The BF compares the likelihood of the findings under H1 relative to the likelihood under the null hypothesis of no effect, H0. In our analysis, H1 is defined as a theoretical distribution assumed to be half normal, with a standard deviation equal to the effect found in the experiment that was the basis for the paradigm we used (Dienes, 2014). Specifically, to test the facial-feedback effect in the camera and no-camera conditions, we assumed the *SD* of the distribution H1 to be the mean difference between the amusement ratings of participants in the lips and teeth conditions (i.e., 0.82) in the original experiment (Strack et al., 1988). For the interaction effect, the *SD* of the distribution of H1 (0.79) was based on the difference between the effect in the no-camera study (Strack et al., 1988) and the effect in the camera study (Wagenmakers et al., 2016).

contrast, in both the original study (Strack et al., 1988) and the replication (Wagenmakers et al., 2016), the DV was the average rating of all four cartoons by each participant. However, there is a difference in the exclusion of participants in the two studies. In the replication project, participants were asked, using a single question, whether they understood the cartoons and were excluded if they answered "no" to this question. In the Strack et al. study, none of the participants was excluded. It could be argued that in the context of evaluating how funny cartoons are, when one indicates that s/he did not understand the cartoon it means that s/he does not find it funny. This raises the possibility that comprehension-based exclusion might differentiate between the outcomes in the original study and the replication. Although we did not preregister this analysis, we believe that the difference between the possible DVs might be informative for future research using this paradigm. Therefore, to examine the role of comprehension-based exclusion, in the analysis presented below we computed the average amusement rating per participant based on all four cartoons, rather than on the basis of the cartoons indicated as being understood (as was done in the main analysis presented above). The analysis is based on the same exclusion criteria as in the main analysis.

A two-way between-participants ANOVA (Camera Presence [yes/no] × Facial Expression [Teeth/Lips]) with planned contrasts was computed on the average amusement rating based on all four cartoons. We also report the Bayes factors of the effects of interest.

As in the previous analysis, the planned contrast analysis revealed a significant facial-feedback effect in the no-camera condition, $t(162) = 2.21$, $p = .028$, $\eta_p^2 = .029$, 95% confidence limits: [.082–1.426], BF = 5.048, such that participants in the teeth condition rated the cartoons as significantly more amusing than did participants in the lips condition ($M_{teeth} = 5.42$, $SD = 1.33$, $M_{lips} = 4.66$, $SD = 1.54$). This effect disappeared in the camera-present condition ($M_{teeth} = 4.73$, $SD = 1.35$, $M_{lips} = 4.91$, $SD = 1.84$), $t(162) = .53$, $p = .60$, $\eta_p^2 = .002$, 95% confidence limits: [−.831–.481], BF = 0.238. Both analyses revealed a significant interaction between the facial expression and camera presence, $t(162) = 1.95$, $p = .026$, one-tailed, $\eta_p^2 = .023$, BF = 3.428.

Discussion

Our results suggest that the original findings of a facial feedback effect (Strack et al., 1988) and the null effect in the replication studies (Wagenmakers et al., 2016) are not contradictory: The facial-feedback effect was present when the camera was absent and participants were not concerned with performance monitoring; The effect diminished when the participants' performance was recorded and participants were warned that their performance was being monitored. Put differently, our findings are consistent with the conjecture that the replication study included a procedural variation (i.e., a presence of a camera) that influenced the effect in question. We would like to emphasize that although it is tempting to interpret our statistical analyses above as showing unambiguously the presence versus absence of the facial-feedback effect, such interpretation should be made with caution. Even a highly diagnostic statistical outcome, such as BF = 8.66, should be interpreted probabilistically, meaning that one should remember that the probability of making the wrong inference is not negligible (10% in the case of BF = 8.66).

Psychology is a cumulative science. As such, no single study can provide the ultimate, final word on any hypothesis or phenomenon. As researchers, we should strive to replicate and/or explicate, and any one study should be considered one step in a long path. In this spirit, let us discuss several possible ways to explain the role that the presence of a camera can have on the facial-feedback effect.

Feeling Observed

The presence (vs. absence) of the camera might have induced the feeling of being observed, which can lead participants to adopt an external perspective of themselves from which their internal information is not available (Hass, 1984; Noah, Schul, & Mayo, in press; Pronin, 2008; Wiekens, 2009). A related theoretical possibility is that participants who feel observed view the self from a third-person perspective (Libby & Eibach, 2011; Libby, Eibach, & Gilovich, 2005; Nigro & Neisser, 1983). Research suggests that compared with participants in a first-person perspective, those who construe an event from a third-person perspective recall less information about the actor's bodily sensations, affective reactions, and psychological states (Libby & Eibach, 2011; McIsaac & Eich, 2002). In terms of our study, it is possible that the presence of the camera led participants to adopt a third-person perspective on the situation, and consequently to become less attuned to the sensory feedback from their facial muscles.

Accountability

Concerns regarding performance monitoring might have increased participants' sense of accountability (Lerner & Tetlock, 2013), which in turn might trigger a stronger tendency to make reason-based judgments. Past research suggests that people who are held accountable are more concerned about information selection and expend greater effort on processing information than do nonaccountable subjects (Rausch & Brauneis, 2015). Using the "fast and the slow" terminology (Kahneman, 2011), it could be argued that accountability shifts decision makers to "slow" processes, perhaps because when they anticipate being accountable, decision makers become more cautious (DeAndrea, Tom Tong, Liang, Levine, & Walther, 2012). Consequently, warning participants about the monitoring of their performance may lead them to be less attentive to the weak signals of their facial feedback and focuses them on the more objective external cues, namely, the actual cartoons.

At this point, we cannot tell which of these mechanisms was responsible for the outcomes. It is quite possible that they worked in tandem, as they are not contradictory. In our past research (Noah, Schul, & Mayo, in press), which tested how the state of mind that might be triggered by feeling observed influences the reliance on metacognitive information in judgment, we attempted to reduce the likelihood of alternative explanations such as accountability through the experimental procedure. However, in the present work we sought to compare the outcomes of the two specific procedures that were used in past studies, one by Strack et al. (1988) and the other by Wagenmakers et al. (2016), and therefore, we repeated the same procedures that were used in those studies.

The Cumulative Nature of Science

The last few years have witnessed a surge of methodological concerns that involve questionable research practices (e.g., Fiedler & Schwarz, 2016; John, Loewenstein, & Prelec, 2012), questionable analytical practices (e.g., Simmons, Nelson, & Simonsohn, 2011), and questionable interpretations of statistics (e.g., Dienes, 2016). There is

no doubt that the field of psychology has made substantial advances in promoting the understanding of the way we should do and report research. However, this has come at a cost. It has introduced a focus on good versus bad researchers rather than on the research (Fiske, 2016). Debates about scientific questions turn into an issue of us-against-them. The unfortunate outcome of this state of affairs is a decrease in cumulative science. Controversies about whose findings are correct almost necessarily preclude a conclusion that both sides might be correct. As Kahneman and Klein (2009) demonstrated, adversarial collaboration can be amply productive. In particular, the strategy of showing that a replication effort can either succeed or fail depending on identified moderators is still rare in the replication literature (Luttrell et al., 2017). Failed replication attempts represent an opportunity to discover new moderators and to test their importance (Van Bavel et al., 2016). Replications can never be exact, and they should be conducted and interpreted in a way that respects the complexity of the psychological phenomena.

Moreover, because the social arena is highly complex it is not surprising that people are sensitive to minute variations in their environment. Thus, failure to capture a theoretically important phenomenon that is created by a specific experimental treatment may stem from participants' reactions to cues that might go unnoticed by researchers. Although context is very hard to study, is it too important to ignore (Van Bavel et al., 2016). Our study illustrates the relevance of this perspective for understanding the failure to replicate the facial-feedback effect, but we see it as only one of many examples for the meaning of failed replications.

We believe that the pendulum has shifted too strongly in the direction of statistical analysis (cf. McShane, Gal, Gelman, Robert, & Tackett, 2017) at the expense of the importance of context. Accordingly, failed replications should trigger not only methodological and statistical discussions, but also theoretically driven analyses of boundary conditions and contextual moderation. This should be followed by testing potential theoretical differences between the original study and failed replication attempt within the same experimental array. In our view, this theoretical and methodological approach would change failed replications from being the "end of the road" regarding entire fields of study and fundamental effects, to a new road for learning and development of our understanding of human nature. In short, the path to understanding social phenomena is long and winding, requiring cooperation, good faith, advanced methodological tools and at the same time, a realization that statistical sophistication cannot replace Lewin's (1943) advice, "there is nothing as practical as a good theory."

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

DeAndrea, D. C., Tom Tong, S., Liang, Y. J., Levine, T. R., & Walther, J. B. (2012). When do people misrepresent themselves to others? The effects of social desirability, ground truth, and accountability on deceptive self-presentations. *Journal of Communication, 62,* 400–417. http://dx.doi.org/10.1111/j.1460-2466.2012.01646.x

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5,* 781. http://dx.doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72,* 78–89. http://dx.doi.org/10.1016/j.jmp.2015.10.003

Eagly, A. H., Wood, W., & Chaiken, S. (1978). Causal inferences about communicators and their effect on opinion change. *Journal of Personality and Social Psychology, 36,* 424–435. http://dx.doi.org/10.1037/0022-3514.36.4.424

Ebersole, C. R., Alaei, R., Atherton, O. E., Bernstein, M. J., Brown, M., Chartier, C. R., . . . Rule, N. O. (2017). Observe, hypothesize, test, repeat: Luttrell, Petty and Xu (2017). demonstrate good science. *Journal of Experimental Social Psychology, 69,* 184–186. http://dx.doi.org/10.1016/j.jesp.2016.12.005

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7,* 45–52. http://dx.doi.org/10.1177/1948550615612150

Fiske, S. (2016, October 31). A call to change science's culture of shaming. [Google Scholar.]. *APS Observer*. Retrieved from https://www.psychologicalscience.org/observer/a-call-to-change-sciences-culture-of-shaming

Forgas, J. P., Brennan, G., Howe, S., Kane, J. F., & Sweet, S. (1980). Audience effects on squash players' performance. *The Journal of Social Psychology, 111,* 41–47. http://dx.doi.org/10.1080/00224545.1980.9924271

Froming, W. J., Walker, G. R., & Lopyan, K. J. (1982). Public and private self-awareness: When personal attitudes conflict with societal expectations. *Journal of Experimental Social Psychology, 18,* 476–487. http://dx.doi.org/10.1016/0022-1031(82)90067-1

Hass, R. G. (1984). Perspective taking and self-awareness: Drawing an *E* on your forehead. *Journal of Personality and Social Psychology, 46,* 788–798. http://dx.doi.org/10.1037/0022-3514.46.4.788

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. http://dx.doi.org/10.1177/0956797611430953

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515–526. http://dx.doi.org/10.1037/a0016755

Laird, J. D., & Lacasse, K. (2014). Bodily influences on emotional feelings: Accumulating evidence and extensions of William James's theory of emotion. *Emotion Review, 6,* 27–34. http://dx.doi.org/10.1177/1754073913494899

Lerner, J. S., & Tetlock, P. E. (2013). Bridging individual, interpersonal, and institutional approaches to judgment and decision making: The impact of accountability on cognitive bias. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 431–457). New York, NY: Cambridge University Press.

Lewin, K. (1943). Psychology and the process of group living. *The Journal of Social Psychology, 17,* 113–131. http://dx.doi.org/10.1080/00224545.1943.9712269

Libby, L. K., & Eibach, R. P. (2011). Visual perspective in mental imagery: A representational tool that functions in judgment, emotion, and self-insight. *Advances in Experimental Social Psychology, 44,* 185–245. http://dx.doi.org/10.1016/B978-0-12-385522-0.00004-4

Libby, L. K., Eibach, R. P., & Gilovich, T. (2005). Here's looking at me: The effect of memory perspective on assessments of personal change. *Journal of Personality and Social Psychology, 88,* 50–62. http://dx.doi.org/10.1037/0022-3514.88.1.50

Luttrell, A., Petty, R. E., & Xu, M. (2017). Replicating and fixing failed replications: The case of need for cognition and argument quality. *Journal of Experimental Social Psychology, 69,* 178–183. http://dx.doi.org/10.1016/j.jesp.2016.09.006

McIsaac, H. K., & Eich, E. (2002). Vantage point in episodic memory. *Psychonomic Bulletin & Review, 9,* 146–150. http://dx.doi.org/10.3758/BF03196271

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2017). Abandon statistical significance. *arXiv Preprint*.

Nigro, G., & Neisser, U. (1983). Point of view in personal memories. *Cognitive Psychology, 15,* 467–482. http://dx.doi.org/10.1016/0010-0285(83)90016-6

Noah, T., Schul, Y., & Mayo, R. (in press). Thinking of oneself as an object of observation reduces reliance on metacognitive information. *Journal of Experimental Psychology: General*.

Pronin, E. (2008). How we see ourselves and how we see others. *Science, 320,* 1177–1180. http://dx.doi.org/10.1126/science.1154199

Rausch, A., & Brauneis, A. (2015). The effect of accountability on management accountants' selection of information. *Review of Managerial Science, 9,* 487–521. http://dx.doi.org/10.1007/s11846-014-0126-8

Scheier, M. F., & Carver, C. S. (1980). Private and public self-attention, resistance to change, and dissonance reduction. *Journal of Personality and Social Psychology, 39,* 390–405. http://dx.doi.org/10.1037/0022-3514.39.3.390

Schlenker, B. R. (1980). *Impression management*. London, UK: Brooks/Cole Publishing Company.

Schwarz, N. (2015). Metacognition. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA handbook of personality and social psychology: Attitudes and social cognition* (pp. 203–229). Washington, DC: American Psychological Association.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Steinmetz, J., Xu, Q., Fishbach, A., & Zhang, Y. (2016). Being observed magnifies action. *Journal of Personality and Social Psychology, 111,* 852–865. http://dx.doi.org/10.1037/pspi0000065

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science, 11,* 929–930. http://dx.doi.org/10.1177/1745691616674460

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54,* 768–777. http://dx.doi.org/10.1037/0022-3514.54.5.768

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *The American Journal of Psychology, 9,* 507–533. http://dx.doi.org/10.2307/1412188

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 113,* 6454–6459. http://dx.doi.org/10.1073/pnas.1521897113

van Bommel, M., van Prooijen, J. W., Elffers, H., & Van Lange, P. A. (2012). Be aware to care: Public self-awareness leads to a reversal of the bystander effect. *Journal of Experimental Social Psychology, 48,* 926–930. http://dx.doi.org/10.1016/j.jesp.2012.02.011

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11,* 917–928. http://dx.doi.org/10.1177/1745691616674458

Wicklund, R. A., & Duval, S. (1971). Opinion change and performance facilitation as a result of objective self-awareness. *Journal of Experimental Social Psychology, 7,* 319–342. http://dx.doi.org/10.1016/0022-1031(71)90032-1

Wiekens, C. J. (2009). *Self-awareness*. Academisch proefschrift. Universiteit van Tilburg.

Zajonc, R. B. (1965). Social facilitation. *Science, 149,* 269–274. http://dx.doi.org/10.1126/science.149.3681.269

Appendix

Analyses According to the Different Exclusion Criteria

The exclusion criteria in the preregistered analysis plan were based on the protocol of Wagenmakers et al. (2016). We departed from the preregistered plan in the following exceptions. First, we did not anticipate in advance that some participants would refuse to hold the pen with their mouth. Those who refused were excluded from the main analysis. Second, we did not anticipate that participants would suspect the cover story of video recording. Those who expressed suspicion about the presence of the camera were excluded as suspicion interferes with reliance on metacognitive feelings (Schwarz, 2015). Lastly, in the preregistration we declared that we plan to replace excluded participants by new ones. However, because we finished collecting the data from 200 participants on the last day of the academic year, we stopped collecting the data at this point, without replacing the excluded participants.

Table 1 reports a series of analyses according to the different exclusion criteria, thus enabling a demonstration of the sensitivity of the outcomes. As in the main analysis in the article, the mean ratings in each condition are based only on cartoons that were reported as comprehended. Table 2 reports the outcomes of the same analyses, this time based on all four cartoons (as in the additional analysis in the article). The results show that the original effect of facial feedback is highly significant in the no-camera condition regardless of the exclusion criteria employed. Analogously, the effect fails to reach significance under the camera condition regardless of the exclusion criteria employed. In fact, this pattern holds even when we include all participants in the analysis (see bottom row).

Table 3 describes the distribution of excluded participants. There is no statistical evidence that exclusion varied systematically as a function of the experimental condition, χ^2 (1, N = 32) = 1.89, p = .169.

(*Appendix continues*)

Table 1
Mean Ratings of Comprehended Cartoons

Criterion	Camera			No camera			Interaction	
	Lips	Teeth	*t* test *p* value	Lips	Teeth	*t* test *p* value	*p* value	*N*
Exclusion based on criteria i–vi (the analysis that is reported in the manuscript)[a]	5.15	5.21	.424	4.92	5.75	.007	.051	166
Exclusion based on criteria ii–vi	5.15	5.21	.426	4.81	5.75	.003	.034	167
Exclusion based on criteria iii–vi	5.15	5.24	.389	4.78	5.75	.003	.033	170
Exclusion based on criteria iv–vi	5.08	5.24	.318	4.77	5.75	.002	.041	172
Exclusion based on criteria v–vi	4.97	5.26	.179	4.80	5.62	.005	.115	192
Exclusion based on criteria i–iv + vi	5.13	5.21	.407	4.92	5.66	.014	.078	170
Exclusion based on criteria i–iv (All preregistered exclusion criteria, no post-hoc criteria)	5.13	5.21	.408	4.92	5.61	.020	.097	172
All participants	4.97	5.26	.174	4.80	5.51	.012	.174	198

Note. The facial feedback effect is not significant in the camera condition according to any of the criteria, and is significant in the no-camera condition according to all of the criteria. The interaction fluctuates according to the exclusion criteria. All *p* values are one-tailed. Because of an error the condition assignment was not recorded for two participants, and therefore they were not included in the analyses.
[a] See exclusion section for details about the criteria.

Table 2
Mean Ratings Based on All Four Cartoons

Criterion	Camera			No camera			Interaction	
	Lips	Teeth	*t* test *p* value	Lips	Teeth	*t* test *p* value	*p* value	*N*
Exclusion based on criteria i–vi (the analysis that is reported in the manuscript)[a]	4.91	4.73	.300	4.66	5.42	.014	.026	166
Exclusion based on criteria ii–vi	4.91	4.73	.304	4.56	5.42	.007	.170	167
Exclusion based on criteria iii–vi	4.91	4.76	.332	4.54	5.42	.006	.017	170
Exclusion based on criteria iv–vi	4.85	4.76	.400	4.52	5.42	.005	.021	172
Exclusion based on criteria v–vi	4.75	4.81	.422	4.57	5.28	.015	.080	192
Exclusion based on criteria i–iv + vi	4.85	4.73	.360	4.66	5.34	.023	.047	170
Exclusion based on criteria i–iv (All preregistered exclusion criteria, no post-hoc criteria)	4.85	4.73	.362	4.66	5.29	.033	.060	172
All participants	4.71	4.81	.369	4.57	5.18	.029	.135	198

Note. The average rating which includes the cartoons that were reported as not understood is lower than the average which includes only the comprehended cartoons. The facial feedback effect is not significant in the camera condition according to any of the criteria, and is significant in the no-camera condition according to all of the criteria. The interaction fluctuates according to the exclusion criteria. All *p* values are one-tailed. Because of an error the condition assignment was not recorded for two participants, and therefore they were not included in the analyses.
[a] See exclusion section for details about the criteria.

Table 3
Distribution of the Excluded Participants

Condition	Lips	Teeth
Camera	8	5
No-camera	7	12

Note. $\chi^2(1, N = 32) = 1.89, p = .169$.

Received November 14, 2017
Revision received February 12, 2018
Accepted February 12, 2018 ■