A multi-semester classroom demonstration yields evidence in support of the facial feedback effect

Abigail A. Marsh, Shawn A. Rhoads, and Rebecca M. Ryan

Georgetown University

Author Note

Abigail A. Marsh, Shawn A. Rhoads, and Rebecca M. Ryan, Department of Psychology, Georgetown University.

Correspondence concerning this article should be addressed to Abigail Marsh, Department of Psychology, Georgetown University, 37th & O Streets NW WGR 302A, Washington, DC 20016

Contact: aam72@georgetown.edu

Abstract

The facial feedback effect refers to the influence of unobtrusive manipulations of facial behavior on emotional outcomes. That manipulations inducing or inhibiting smiling can shape positive affect and evaluations is a staple of undergraduate psychology curricula and supports theories of embodied emotion. Thus, the results of a Registered Replication Report indicating minimal evidence to support the facial feedback effect were widely viewed as cause for concern regarding the reliability of this effect. However, it has been suggested that features of the design of the replication studies may have influenced the study results. Relevant to these concerns are experimental facial feedback data collected from over 400 undergraduates over the course of 9 semesters. Circumstances of data collection met several criteria broadly recommended for testing the effect, including limited prior exposure to the facial feedback hypothesis, conditions minimally likely to induce self-focused attention, and the use of moderately funny contemporary cartoons as stimuli. Results yielded robust evidence in favor of the facial feedback hypothesis. Cartoons that participants evaluated while holding a pen or pencil in their teeth (smiling induction) were rated as funnier than cartoons they evaluated while holding a pen or pencil in their lips (smiling inhibition). The magnitude of the effect overlapped with original reports. Findings demonstrate that the facial feedback effect can be successfully replicated in a classroom setting and are in line with theories of emotional embodiment, according to which internal emotional states and relevant external emotional behaviors exert mutual influence on one another.

Keywords: Facial feedback, embodied emotion, facial expression, humor, positive affect

A multi-semester classroom demonstration yields evidence in support of the facial feedback effect

According to the *facial feedback hypothesis,* external facial behavior can influence internal emotional experiences (McIntosh, 1996). Smiling, according to this hypothesis, is not simply an outward display signifying a positive emotional state but can also cause increases in positive emotional states like happiness and amusement. In a paradigm that has now been cited over 1,700 times and is widely incorporated into introductory psychology curriculum, Strack, Martin, and Stepper (1988) demonstrated that an unobtrusive manipulation of facial behavior, such as asking participants to hold a pen in their teeth (which facilitates smiling) causes them to rate cartoons as more humorous than when they hold a pen in their lips (which inhibits smiling). Variations of this effect have been demonstrated in dozens of subsequent studies (Coles, Larsen, & Lench, 2017; Strack, 2016) and are consistent with theories of emotional embodiment, whereby internal emotional states and relevant external emotional behaviors exert mutual influence on one another (Laird & Lacasse, 2014; Price & Harmon-Jones, 2015; Winkielman, Niedenthal, Wielgosz, Eelen, & Kavanagh, 2015).

However, a recent and highly publicized Registered Replication Report (RRR) in a large sample failed to replicate the facial feedback effect (Wagenmakers et al., 2016). Following a meta-analysis of data collected by 17 participating laboratories employing similar paradigms as in the original experiment by Strack and colleagues, the authors reported a facial feedback effect size of 0.03, which was substantially smaller than the originally reported effect sizes of 0.82. The 95% meta-analytic confidence interval reported by Wagenmakers and colleagues ranged from −0.11 to 0.16, which overlapped with zero. This failed replication received significant scholarly

attention and was described as symptomatic of a larger crisis of replication in psychology (Skibba, 2016).

In a response, Strack (2016) praised the researchers' goals but highlighted several features of the RRR that deviated from the original paradigm and may have influenced the observed effect. One was the use of video cameras to record participants' facial movements during testing. Video recording was important for confirming consistency across participating labs. However, being recorded can increase self-focused attention and change attitudinal reports, evaluations of material presented during recording, and the desire to maintain attitude-behavior consistency (Wicklund & Duval, 1971; Davis & Brock, 1975) and was recently demonstrated to reduce the facial feedback effect (Noah, Schul, & Mayo, 2018). A second concern was the recruitment of participants from psychology subject pools. Because the facial feedback effect is discussed in many psychology textbooks, this recruitment strategy introduced the possibility that students were previously aware of the effect. Although the authors eliminated participants who correctly guessed the study hypotheses, when effect sizes among students recruited from psychology subject pools versus other sources were compared, smaller effect sizes were observed in psychology subject pool samples. This effect significantly deviated from zero, $t(2) = 5.09$, $p = .037$, in the direction of the original result (Strack, 2016). A third concern was the use of *Far Side* cartoons from the 1980s that may be less relevant to contemporary students and which resulted in exclusion of some participants who did not understand the cartoons.

Prior to and during time of the RRR, experimental facial feedback data were being collected continuously for instructional purposes in a large undergraduate introductory psychology class. These data were collected in such a way that the effect of facial feedback could be statistically evaluated. The nature of the testing circumstances also met several key criteria broadly

recommended for testing the effect (Strack, 2016, 2017; Wagenmakers et al, 2016). First, data were collected in a large classroom setting, precluding individualized interaction with the instructor (the experimenter) during testing and mitigating potential experimenter effects and self-focused attention. Second, testing was always performed in the classroom two weeks prior to coverage of emotion or the facial feedback effect, ensuring participants had minimal formal exposure to the effect. Third, contemporary Dilbert cartoons were used (© 2008) that were rated to be moderately funny, minimizing misunderstandings and the potential for ceiling or floor effects. And fourth, the total sample was sufficiently large to test the effect (>400 students).

Although data were initially collected for instructional purposes, they were subsequently analyzed to assess whether they were consistent with the facial feedback hypothesis.

Method

Participants

Over nine semesters beginning in the Fall of 2008, 446 male and female undergraduate students taking an introductory psychology course at Georgetown University participated in a facial feedback experiment during a lecture. This sample included all students in attendance during each day of testing. Data were submitted anonymously without identifying details, including age, gender, education level, or socioeconomic status. The methods and data collected were determined to be exempt from IRB review by the Georgetown University Institutional Review Board.

The task was introduced early in a lecture on learning, which was the 7th lecture of the semester and the last topic introduced before the first midterm examination. This lecture took place at least two weeks prior to the lecture on emotion and to students being assigned the accompanying textbook chapter (Schacter, Gilbert, Nock & Wegner, 2009) that covered the

facial feedback effect. Students with AP or IB psychology credit were exempt from the course. At the time of the experiment, participants therefore had minimal prior formal exposure to the facial feedback effect.

Facial feedback task

The experimental manipulation was consistent with the original facial feedback paradigm (Strack, Martin, & Stepper, 1988, Study 1). After the topic of associations in classical conditioning were introduced, students were instructed to take out a pen, pencil, or similarly shaped object. The instructor then divided students in the class in half according to their seating position (to the left or right of the center of the classroom). Students were asked to write "Right" or "Left" on their paper to indicate where in the class they were seated. Students on the right side of the room were verbally instructed to hold the pen in their teeth without letting their lips touch it, and students on the left were instructed to hold the pen in their lips without letting their teeth touch it. Images from Strack and colleagues (1988) were projected on the screen to illustrate how students on each side of the class should position their pen or pencil (Figure 1). The instructor also demonstrated the correct techniques. Other than the image, no details of the original study or its authors were provided.

While holding their pens in their mouth, students were shown a three-panel Dilbert cartoon (Cartoon 1, http://dilbert.com/strip/2008-09-17) via the overhead projector and were asked to evaluate how funny it was on a 7-point scale (1 = not at all funny, 7 = extremely funny), rather than the 0-9 scale used by Strack and Wagenmakers. Students were instructed to decide on their rating while holding their pens in their mouths, and then afterward to remove the pens and write their evaluation of the cartoon next to the designation "C-1".

Next, students were asked to switch their pen to the opposite configuration. For clarity, a second image was presented with the illustrations of correct pen positions for the two sides of the classroom. Students were then asked to evaluate a second three-panel Dilbert cartoon (Cartoon 2, http://dilbert.com/strip/2008-09-12) following the same procedure. (The same cartoons were used every semester.) After students rated the second cartoon, the nature of the experiment and the facial feedback hypotheses were explained. At the end of class, students brought their ratings to the front of the classroom. No identifying information was collected. Students were not required to participate and received no credit for providing their ratings. Scores were collected and the data entered by a graduate teaching assistant.

This procedure resulted in a counterbalanced design such that approximately half of participants ($N = 226$) evaluated the Cartoon 1 with pen-in-teeth then Cartoon 2 with pen-in-lips (Sequence 1), whereas the other half of participants ($N = 220$) rated Cartoon 2 with pen-in-teeth then Cartoon 1 with pen-in-lips (Sequence 2).

Statistical analysis

A total of 856 observations were available for statistical analyses (428 participants × 2 ratings; see Supplementary Materials for raw data). Listwise deletion was performed for 8 students with missing data and 10 students with at least one rating inconsistent with task instructions, such as values outside the 1-7 range or non-integer values). To assess whether mean differences in amusement ratings were significant across facial feedback conditions, within-subject amusement ratings were analyzed in a mixed-effects hierarchical linear model with participants nested within classes. The 3-level hierarchical linear model controlled for variation in amusement ratings across classes and the variation in amusement ratings across participants within classes by including random intercepts for class and for participants within class,

respectively. The categorical variable of facial feedback condition was included as a within-subject fixed slope and the sequence of cartoons was included as a between-subject fixed-effect covariate of no interest.

Results

An independent sample of 97 adults who were recruited through Mturk and selected to be comparable in age to our student population (specified age range 18-25, *M* age = 23.33, *SD* = 1.64; 3 participants were excluded whose ages were outside the specified range), and who included 56 females, 40 males, and 1 who did not specify their gender, evaluated both cartoons on a 7-point scale (1 = not at all funny, 7 = extremely funny) and rated both cartoons as moderately funny, (Cartoon 1 $M = 3.85$, $SD = 1.38$; Cartoon 2 $M = 4.11$, $SD = 1.65$), $t(96) = 1.41$, $p = 0.162$, $CI_{95\%}$ [-0.11, 0.65]. Students also rated both cartoons as moderately funny, although their mean ratings were lower, with Cartoon 2 being funnier than Cartoon 1 (Cartoon 1 $M = 2.74$, $SD = 1.39$; Cartoon 2 $M = 3.24$, $SD = 1.48$), $t(432)=5.56$, $p < .001$ $CI_{95\%}$ [0.33 .68]. Main effects and interactions of students' mean amusement ratings were then compared across participants, nine classes, two sequence groups, and two facial feedback conditions. The average amusement ratings between sequences did not significantly differ. Amusement ratings for Sequence 1 (Cartoon 1, Cartoon 2) were 0.16 points lower than Sequence 2 (Cartoon 2, Cartoon 1) ($\beta = -0.20$, $SE = 0.10$, $Z = -1.94$, $p = 0.052$, $CI_{95\%}$ [-0.40, 0.02]).

More importantly, the mean difference in amusement rating as a function of facial feedback condition across individuals within classes, controlling for sequence, was statistically significant ($\beta = 0.40$, $SE = 0.09$, $Z = 4.33$, $p < 0.001$) with cartoons evaluated with pen-in-teeth rated as funnier ($M = 3.18$, $SD = 1.46$) than cartoons evaluated with pen-in-lips ($M = 2.78$, $SD = 1.42$). On average, amusement ratings for the pen-in-teeth condition were 0.40 points higher than

the pen-in-lips condition, with a small-to-medium effect size, $d = 0.28$, $CI_{95\%}$ [0.22, 0.58]. The observed 0.40 unit mean difference on the 7-point scale is equivalent to a difference of 0.57 units on the 10-point scale used by Strack (1988) and Wagenmakers (2016). The $CI_{95\%}$ [0.31, 0.83] around this adjusted value overlaps with the mean effect size (0.82) originally reported by Strack (1988). The consistency of the effect can be observed in Figure 2b, in which mean ratings by condition are plotted for each semester. All values but one lie above the main diagonal, indicating reliably higher ratings for pen-in-teeth, as predicted by the facial feedback hypothesis. Variation in ratings attributable to variance across classes was small (ICC=1.72%) and the variation in ratings attributable to variance between students within classes was moderate (ICC=12.39%).

Analyses were repeated following pairwise deletion of only the 9 missing data entries, with all other values retained. These analyses included a total of 883 observations available for statistical analyses (445 participants × 1 pen-in-teeth rating, 438 participants × 1 pen-in-lips rating). No changes in outcomes were observed. Average amusement ratings between sequences did not significantly differ. The ratings for Sequence 1 (Cartoon 1, Cartoon 2) were 0.16 points smaller than Sequence 2 (Cartoon 2, Cartoon 1) ($\beta = -0.16$, $SE = 0.10$, $Z = -1.60$, $p = 0.11$, $CI_{95\%}$ [-0.36 0.04]). Amusement rating between facial feedback conditions across individuals within classes, controlling for sequence, again significantly differed (β=0.42, $SE = 0.09$, $Z = 4.57$, $p < 0.001$, with amusement ratings for the pen-in-teeth condition being 0.42 points higher than the pen-in-lips condition, $d = 0.30$, $CI_{95\%}$ [.24, .61]. Variation in ratings attributable to variance across classes was again small (ICC = 1.34%) and variation in ratings attributable to variance between students within classes was moderate (ICC = 9.95%).

Finally, we sought to replicate the between-subjects effect identified by Strack and colleagues (1988) by conducting group comparisons for the first trial only (Cartoon 1 pen-in-teeth versus pen-in-lips) in a hierarchical linear model with participants nested within classes and 442 available observations (4 participants with missing data for Cartoon 1 were excluded listwise). Results again identified a significant mean difference in ratings as a function of feedback condition across groups within classes ($\beta = 0.59$, $SE = 0.13$, $Z = 6.61$, $p < 0.001$), with students in the pen-in-teeth condition rating Cartoon 1 as funnier ($M = 3.06$, $SD = 1.45$) than student in the pen-in-lips condition ($M = 2.46$, $SD = 1.28$). Average amusement ratings for the pen-in-teeth condition were 0.59 points higher than the pen-in-lips condition, with a medium effect size, $d = 0.44$, $CI_{95\%}$ [.34, .84].

Discussion

The results of this study replicate the facial feedback effect in the type of classroom setting in which this effect is often taught. In a large sample of undergraduate students beginning an introductory psychology course, cartoons evaluated during a manipulation that simulates smiling (holding a pen in the teeth) were rated as more humorous than when the cartoons were evaluated during a manipulation that inhibits smiling (holding a pen in the lips). These results were obtained following an analysis plan selected based on clustering of ratings among classes. The analyses indicated that the manipulation resulted in a small-to-medium effect size, the magnitude of which overlapped with the effect observed in the original report of the facial feedback effect.

The experimental conditions featured several strengths that may have contributed to the observed effect. The nature of the testing setting precluded experimenter effects related to differential treatment by condition, as both conditions were run simultaneously. It also minimized the likelihood of previous formal exposure to the facial feedback effect, as the

experiment was conducted in introductory psychology students several weeks before the textbook chapter describing the effect was assigned (it seems safe to assume no students read several chapters ahead). It is, however, important to note we do not have formal confirmation of participants' prior lack of exposure to the feedback effect. Participants were not video recorded, minimizing self-focused attention, which can alter response styles, affective experiences, and self-report motivations. And contemporary cartoons rated as moderately funny were used as stimuli.

The paradigm diverged from the original facial feedback experiment in several respects. They include the classroom setting in which testing was conducted; the fact that each participant rated two cartoons rather than four; the fact that it featured a within-subjects rather than between-subjects design; the absence of a cover story about piloting a study for future research regarding populations with disabilities to explain the manipulation; the use of a 7-point scale rather than a 10-point scale; the fact that the experiment was part of a classroom lecture about learning (specifically, about the acquisition of conditioned associations) rather than following a line-drawing task; the fact that correct positioning of pens could be monitored only within the limits of a group setting; the fact that participants selected but did not write down their ratings with their pens in their mouths; and the lack of individualized follow-up with participants regarding their beliefs about the experiment, precluding exclusion of participants for suspicions regarding the study goals. (It is notable, however, that when the instructor presented students with their results in the ensuing class, the most commonly verbalized reaction was surprise or disbelief that the manipulation could have possibly affected their ratings.)

Results of two recent papers (Coles, Larsen, & Lench, 2017; Noah, Schul, & Mayo, 2018) found that the facial feedback effect can be moderated by various factors. Noah and colleagues

found that the effect can be reduced by video-recording participants. The meta-analysis by Coles and colleagues determined that another moderator is the choice of question, with evaluations of the stimulus quality (i.e., how funny the cartoon is) showing larger effect sizes, but also more evidence of publication bias, than ratings of amusement—although fewer estimates of the effect on stimulus quality were available for analysis. The most important moderator that could be accounted for statistically in the meta-analysis was the specific stimuli that were used during testing. The choice of moderately funny contemporary cartoons in the present study may have contributed to the effects we observed, as may other variables that have not been identified. The consistency of the observed effect with original reports despite methodological differences, however, could be interpreted in support of the effect's robustness.

Overall, the results of the study are consistent with the notion that unobtrusive manipulations of facial behavior can reliably shape emotional experiences and outcomes, in line with theories of emotional embodiment.

References

Coles, N. A., Larsen, J. T., & Lench, H. C. (2017). A meta-analysis of the facial feedback hypothesis literature. *Open Science Framework*. https://osf.io/ahcs8/

Davis, D. & Brock, T. C. (1975) Use of first person pronouns as a function of increased objective self-awareness and performance feedback. *Journal of Experimental Social Psychology, 11,* 381-388. https://doi.org/10.1016/0022-1031(75)90017-7

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

McIntosh, D. (1996). Facial feedback hypotheses: Evidence, implications, and directions. *Motivation & Emotion, 20,* 121–147.

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality & Social Psychology, 114,* 657-664.

Price, T. F. & Harmon-Jones, E. (2015). Embodied emotion: the influence of manipulated facial and bodily states on emotive responses. *WIReS Cognitive Science, 6,* 461-473. https://doi.org/10.1002/wcs.1370

Skibba, R. (2016). Psychologists argue about whether smiling makes cartoons funnier. *Nature News,* November 3, doi:10.1038/nature.2016.20929.

Strack, F. (2016). Reflections on the smiling registered replication report. *Perspectives on Psychological Science, 11,* 929-930. https://doi.org/10.1177/1745691616674460

Strack, F. (2017). From data to truth in psychological science. A personal perspective. *Frontiers in Psychology,* 00702. https://doi.org/10.3389/fpsyg.2017.00702

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 & Social Psychology, 54,* 768-777. doi:10.1037/0022-3514.54.5.768

Schacter, D. L., Gilbert, D. T., Nock, M. K., & Wegner, D. M. (2009). *Psychology* (1st Ed). New York: Worth.

Wagenmakers, E.-J., Beek, T., Dijkhoff, T., Gronau, Q. F., Acosta, A., et al. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science,* 11, 917–928. https://doi.org/10.1177/1745691616674458

Winkielman, P., Niedenthal, P., Wielgosz, J., Eelen, J., & Kavanagh, L. C. (2015). Embodiment of cognition and emotion. In M. Mikulincer & P. R. Shaver (Eds.) *APA Handbook of Personality & Social Psychology* (pp. 151-175). Washington, DC: American Psychological Association.

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. https://doi.org/10.1016/0022-1031(71)90032-1

Acknowledgements

Thanks to Kristin Brethel-Haurwitz, Evan Gordon, Caitlin Hines, Elisabeth McClure, Marisa Putnam, Sylvia Rusnak, Jessica Simon, and Sarah Vidal for assistance with data collection and entry.

Figure Captions

Figure 1. Screen capture of illustration used to illustrate correct pen positions to students (from Strack et al., 1988).

Figure 2. a) Mean scores (SD) for each condition by semester; b) Ratings by semester for the pen-in-lips condition is plotted against that of the pen-in-teeth condition. Values consistently lie above the main diagonal, as predicted by the facial feedback hypothesis.

Figure 1

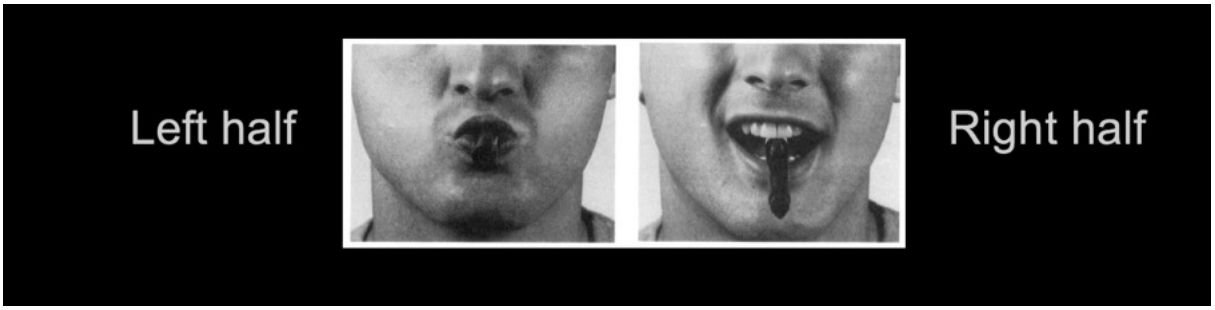

Figure 2