

Methodological Challenges for Internal Validity in Facial Feedback Research: A Replication Study

Bachelor Degree Project in Cognitive
Neuroscience

First Cycle 22.5 credits

Spring term Year 23-24

Student: Conny Mikszath

Supervisor: Andreas Kalckert

Examiner: Paavo Pylkkänen

Abstract

This replication study addresses the methodological challenges impacting the internal validity of research on the Facial Feedback Hypothesis (FFH), which proposes that facial expressions can influence emotional experiences. Recent studies have yielded inconsistent results and unexplained data, prompting an investigation of potential methodological explanations for the variability observed in a recent study by Coles et al. (2022). The current study partially replicated the original study with an added set of questions aimed at further understanding the unexplained data. A total of 67 participants were recruited, all of whom were tested in person, whereas the original study was conducted entirely online. The replication did not yield a statistically significant facial feedback effect and did not indicate a reduction in extreme response patterns relative to the original study. Both quantitative and qualitative measures were employed to assess differences in rating emotions while posing smiling or blank facial expressions. The current findings suggest that there are likely several confounding factors that complicate any interpretation of the original results. Some participants reported that while they were performing the mimicry task, which is expected to induce the facial feedback effect, they had been thinking about how they had felt happy in the past. Without an understanding of the extent or influence of these internal thoughts, these reports raise questions regarding the study's internal validity. This study contributes to the ongoing dialogue regarding the replicability of the FFH and advocates for more critical thinking and rigorous investigations regarding potential confounds when measuring the effect of facial expressions on emotional experiences.

Keywords: facial feedback hypothesis, replication study, internal validity

Methodological Challenges for Internal Validity in Facial Feedback Research

Happiness is highly valued by people across the world, making it an area well worth researching (Diener, 2000). As a psychological construct, it has historically been defined in varied ways, often composed of multiple distinct components (Kringelbach & Berridge, 2010). Diener (2000) describes it as a colloquial term for subjective well-being, which he defines as comprising three elements: high life satisfaction, frequent positive affect, and infrequent negative affect. Seligman (2011), on the other hand, argues for a broader conceptualization that comprises five elements: positive emotion, engagement, relationships, meaning, and a sense of accomplishment. While all of Seligman's five elements are reflected in some laypeople's conceptions of happiness, Delle Fave et al. (2011) found substantial variation in how individuals define the term, with harmony and balance being the most commonly reported themes. In other words, we must be wary of the possibility that we are not measuring the same construct across all individuals when asking them to rate their happiness.

This variation in definitions, combined with the inherently subjective nature of happiness, makes it a particularly challenging concept to measure (Kringelbach & Berridge, 2010). In addition, there is also the risk of p-hacking, where studies are conducted in flexible enough ways to produce significant results that can then be selectively reported (Folk & Dunn, 2024). To mitigate this issue, researchers have increasingly adopted pre-registration, where they can publicly document the details of their research plans, such as hypotheses, methods, and analysis strategies, thereby restricting later publications to those pre-specified procedures. Folk and Dunn (2024) conducted a systematic review that included only pre-registered experiments specifically investigating methods for increasing happiness. Although the effect sizes of the reviewed interventions were generally modest, the facial feedback intervention tested in Coles et al. (2022) was among those with the largest effects. The possibility that facial feedback interventions could be used to improve well-being was supported by some of the authors, but any attempt to apply such interventions meaningfully would require a clearer understanding of the mechanisms involved.

The intervention in Coles et al. (2022) is based on the facial feedback hypothesis (FFH), a theoretical framework that proposes facial expressions can influence emotional experience. This hypothesis was inspired by ideas from Darwin (1872), who noted that emotional experiences could be either intensified or diminished if one freely expressed the emotion physically or suppressed it, respectively, and William James (1884), who suggested that physical states not only modulate emotions but are what generate them. Indeed, major theories of emotion commonly include expression as a component of emotion, along with autonomic responses, action tendencies, appraisals, and feelings (Sander, 2013). The extent to which components like expression, physiology, and subjective experience align with each other, a concept termed emotional coherence, has been debated. Mauss et al. (2005) provided evidence for significant associations among these systems, though the degree of coherence varied. Ekman (1992) similarly argued that coherence may depend on individual factors, suggesting that chronic suppression of emotional expression or feeling can disrupt the coherence between expressive behavior and autonomic responses.

Looking at the history of facial feedback research, it is possible that we have already observed the kind of disruption Ekman (1992) described, though perhaps in an inverted form, where emotion is expressed without being felt. The most well-known early study, conducted by Strack et al. (1988), found that participants who held a pen in their mouth in a way that facilitated a smiling facial expression without their awareness rated cartoons as funnier than those whose task interfered with the ability to form a smile. However, subsequent investigations yielded mixed results. A high-powered multi-lab replication by Wagenmakers et al. (2016), involving 17 laboratories and nearly 1,900 participants, did not find evidence for the original effect. Strack (2016) later argued that the presence of a camera in the replication may have interfered with participants' perception of their internal experiences, perhaps by making them more self-aware or altering how they evaluated their own emotions. Supporting this possibility, Noah et al. (2018) replicated the original study and found that the facial feedback effect was present in the absence of a camera but disappeared when a camera was present.

One interpretation of this finding is that the camera context itself may have disrupted the expected association between smiling and positive affect. Cameras are commonly associated with posed smiling, a context in which people often smile even when they are not experiencing genuine positive emotion. Over time, such repeated expression without feeling might potentially weaken the emotional significance of smiling in these contexts, particularly among individuals who feel uncomfortable being photographed. While Ekman (1992) emphasized that inhibiting expression or feeling can disrupt coherence, the camera context may reflect a different kind of interference: repeated expression in the absence of feeling, especially in a highly scripted and socially performative setting. If the facial feedback effect relies on associative mechanisms such as classical conditioning, then it would make sense that coherence is not globally lost but rather contextually disrupted, a pattern consistent with how conditioned responses can be suppressed or overridden in specific contexts (Bouton, 2002; Harris et al., 2000).

The idea that facial feedback effects may be modulated by conditioning mechanisms has been discussed by several researchers (e.g., Buck, 1980; Kleck et al., 1976). Borg et al. (2016) demonstrated that both facial expressions and emotional valence can be independently conditioned and extinguished. However, their study did not test whether facial expressions themselves could, through conditioning, come to elicit or modulate emotional states. By combining the idea that peripheral feedback can generate conditioned internal reactions (Buck, 1980) with appraisal theories of emotion, which treat interoceptive signals as evidential inputs (Sander, 2013), we may hypothesize a pathway from facial expression to experienced emotion, partly moderated by conditioning.

If such mechanisms exist, it would suggest that facial feedback effects vary across individuals depending on the extent to which such conditioning has been reinforced or extinguished. Consistent with this possibility, Coles et al. (2022) observed substantial individual differences in happiness ratings across facial expression conditions in their large-scale replication. As visualized in Appendix C, nearly a quarter of participants reported lower

happiness, and only slightly more than half reported higher happiness, when posing a smile compared to a blank expression. If facial feedback interventions are to be seriously considered as strategies for improving well-being, it is essential to understand the mechanisms that underlie this wide variability in affective outcomes.

Investigating and Exploring Limitations in Coles et al. (2022)

Anonymity and Trolling

Coles et al. (2022) conducted a high-powered multi-lab investigation of the facial feedback hypothesis, modeled on and extending the original study by Strack et al. (1988). Their design included several theoretically relevant variables, such as participant awareness of the hypothesis and multiple indicators of pose quality, including self-reported compliance with the instructions, similarity to an exemplar expression, and perceived genuineness of the posed expression. None of these variables appear sufficient to explain the negative effects observed in nearly a quarter of participants, especially those cases where participants reported minimal happiness while smiling and near-maximum happiness while posing a blank expression. Some of these extreme responses may have resulted from participants emboldened by the anonymity of the online setting to respond deceitfully or unseriously. This risk is supported by findings from Nitschinsk et al. (2022), who showed that anonymity increases the likelihood of trolling in online environments. To assess whether such extreme responses could be reduced, the present study was conducted in person.

Conditioning-Based Variability

Another explanation worth investigating, which may account for some of this variability, is the conditioning-based account previously discussed. If facial feedback effects are to some extent modulated by conditioning mechanisms (e.g., Hofmann et al., 2010; De Houwer, 2020), then individuals who have frequently smiled without experiencing genuine positive affect, or while feeling negative affect, may develop learned associations that weaken or even reverse the commonly discussed link between smiling and happiness. In such cases,

the act of producing a smiling expression may not evoke the expected increase in happiness and could instead result in reduced or even negative affective ratings (Kleck et al., 1976). Whether conditioned associations could cause this type of variability remains a gap left unaddressed by Borg et al. (2016), and is therefore explicitly assessed here. To examine this possibility, the present study used self-reported frequency of expressing happiness through facial expressions without feeling happy as a potential moderator. This was assessed using adapted items from the happiness subscale of the Discrete Emotions Emotional Labor Scale (DEELS; Glomb & Tews, 2004).

Cognitive Strategies and Affective Decision-Making

As noted earlier, major theories of emotion include appraisal as a core component of emotional experience, alongside expression, autonomic response, action tendencies, and subjective feeling (Sander, 2013). While appraisal is traditionally understood as the evaluation of internal and external cues to determine emotional meaning, recent models have extended this view by framing appraisal as a dynamic, integrative process. In particular, Teoh et al. (2023) propose that emotion self-reports should be understood not as passive readouts of felt states, but as outcomes of affective decision-making where interoceptive signals, proprioceptive feedback, contextual appraisals, and action tendencies are integrated and weighted to construct a judgment. Under this framework, variability in reported emotional experience may reflect not only differences in affect but also differences in how individuals process and interpret the inputs that guide their emotional decisions. For example, Robinson and Clore (2002b) demonstrated that emotion ratings can be influenced by the structure and time frame of the task, with participants showing systematic differences in both response latency and reported intensity depending on whether judgments were based on episodic memory or semantic self-knowledge. While their study manipulated time frame explicitly, the present study entertains the possibility that such strategic shifts might also be implicitly induced by facial expression. One speculative possibility is that posing a blank expression provides little affective information to anchor judgment, leading participants to default to

semantic self-knowledge, whereas a smiling expression may serve as a salient cue that prompts interpretation through episodic recall. If facial expressions systematically influence the type of cognitive strategy employed, this could create bias in affective self-reports even in the absence of genuine emotional change. Without relying on a specific theoretical framework, the present study included a set of exploratory post-task questions aimed at identifying patterns and discrepancies in how participants constructed their emotion ratings. These questions were not designed to test any particular hypothesis, but rather to uncover potentially systematic biases in judgment strategies and patterns that could inform the development of more precise, theory-driven hypotheses for future research.

Unaddressed Order Effects

One additional limitation relevant to the present study concerns Coles et al.'s (2022) treatment of order effects. In their design, the first and last blocks were fixed, while the order of the two central conditions, blank and mimic, was counterbalanced across participants. However, they did not assess whether this counterbalancing effectively mitigated any order effects. This reflects a seemingly common misconception in experimental design: that counterbalancing alone eliminates the risk of order effects. As emphasized by Wänke and Schwarz (1997), counterbalancing may reduce some systematic biases, but it does not prevent all residual order-related effects, particularly in within-subjects designs, where participants make repeated judgments that may be influenced by earlier responses. For example, anchoring effects can occur when ratings in one condition establish a reference point that unintentionally shapes responses in the next. Without explicitly testing for order effects, researchers cannot determine the extent to which observed outcomes are attributable to the intended manipulation versus variation introduced by condition order. Ruling out order effects is essential, as their presence would compromise the internal validity of facial feedback research. Although no specific predictions were made in advance and no theoretical account was assumed, the condition order for each participant was recorded and included in an exploratory analysis to test for potential order effects.

Overview of the Present Study

In light of the theoretical concerns and methodological gaps outlined above, the present study aimed to examine the reliability and internal validity of the facial feedback effect through a partial in-person replication of Coles et al. (2022). Due to time constraints, the study replicated only one of the six experimental conditions used in the original design, which featured three types of facial movement tasks (facial mimicry, pen-in-mouth, and voluntary facial action), each crossed with the presence or absence of positive visual stimuli. Focusing exclusively on the facial mimicry condition without stimuli allowed for greater statistical power within that condition while still supporting a meaningful investigation of potential mechanisms underlying the variability observed in prior work. The procedure for this condition was kept identical to that used in the original study, except that the experiment was conducted in person rather than online. Upon completion of the experimental procedures, participants completed a set of post-experiment survey questions. These included all items from the original study except for the body awareness questions, which were replaced by two new sets of questions intended to extend the scope of the original investigation (see Appendices A and B for full item sets).

Hypotheses

- H1. *Facial feedback effect*: Participants will, on average, report higher levels of self-reported happiness (assessed via the composite of happiness, enjoyment, satisfaction, and liking) after posing a smiling expression compared to a blank expression.
- H2. *Trolling hypothesis (extreme differences)*: The proportion of participants reporting unusually large affective differences between blank and mimicry conditions will be smaller in the present in-person study compared to the online sample reported in Coles et al. (2022).

- H3. *Conditioning hypothesis (emotional labor)*: Higher self-reported frequency of smiling without feeling happy, assessed via an adapted DEELS item, will predict a smaller or negative facial feedback effect.
- H4. *Hypothesis awareness*: Participants who indicate they were aware of the hypothesis will show a larger facial feedback effect than those who were unaware.
- H5. *Compliance with expression instructions*: Higher self-reported compliance with expression instructions will be associated with a larger facial feedback effect.
- H6. *Similarity to exemplar*: Greater perceived similarity between the participant's expression and the target exemplar image will be associated with a larger facial feedback effect.
- H7. *Perceived genuineness*: Higher perceived genuineness of one's own posed expression will be associated with a larger facial feedback effect.

Methods

Participants

By estimating an average of five to six participants per day, the time period for collecting data was set to two weeks, after which data collection would end. Forty participants were recruited by in-person solicitation near the rooms on campus used for the experiment, and twenty-seven participants were recruited via personal networks (family, friends, neighbors, and work colleagues). Two participants aborted the experiment because they suspected that they had misunderstood the tasks and were not included in the analysis. A total of 65 individuals completed the study, including 40 men (61.5%) and 25 women (38.5%), with a mean age of 29.60 years ($SD = 10.78$). Of these, 24 participants passed all predefined inclusion criteria, including successful completion of both attention checks, posture quality ratings, and reporting unawareness of the study's hypothesis. In line with the original study, 98% of participants reported at least some compliance with the facial

expression instructions. Furthermore, 69% passed both attention checks, 92% passed the similarity rating, and 68% passed the awareness criterion.

Power Analysis and Sample Size Determination

To estimate the required sample size, an effect size (Cohen's d_z) was derived from the original dataset for the mimicry/no-stimuli condition in Coles et al. (2022), which was the condition replicated in the present study. Specifically, 10,000 bootstrap iterations were conducted on the paired-condition data using the Wilcoxon signed-rank test for matched pairs. This provided a robust estimate of the within-subjects effect size (d_z) appropriate for the statistical test used in the current analysis. The value of d_z was used rather than Cohen's d because the design involves dependent samples, and G*Power requires d_z for power calculations involving within-subjects tests such as the Wilcoxon signed-rank test for matched pairs.

The bootstrap procedure yielded an effect size estimate of $d_z = 0.506$. This value was entered into G*Power (version 3.1.9.7), specifying a Wilcoxon signed-rank test for matched pairs, with $\alpha = 0.05$ and power = 0.80. The resulting recommended sample size was 27 participants. To validate this estimate under the assumptions of the non-parametric test, 10,000 bootstrap simulations of the Wilcoxon signed-rank test were conducted using the same effect size. These simulations indicated that 29 participants would be required to reach the desired power of 0.80. Estimating the calculated awareness pass rate of 48.8% in Coles et al. (2022) to be similar in the present study, the total sample size target was set to 60 to ensure adequate statistical power after exclusions.

As shown in Appendix C, the distribution of affective difference scores in the original data is markedly non-normal, with a high frequency of zero differences and an asymmetric, potentially multi-peaked shape. This violates the assumptions of the paired-sample t-test and supports the decision to use the Wilcoxon signed-rank test in the current replication.

Design

The study used a within-subjects experimental design, replicating the mimicry/no-stimuli condition from Coles et al. (2022). Each participant completed four facial movement tasks in a fixed sequence. The first and fourth tasks involved non-expressive movements designed to serve as cover tasks and obscure the study's true purpose. The second and third tasks, which involved mimicking a smiling expression posed by four actors in a picture collage and maintaining a blank facial expression, constituted the levels of the independent variable. The order of these two tasks was assigned using a biased coin design (BCD; Efron, 1971), in which the probability of assignment was temporarily adjusted to favor the underrepresented order. This approach preserved random assignment while reducing the risk of substantial imbalance in condition order.

Participants were informed via a cover story that the study investigated the effects of physical movements and cognitive distractions on mathematical problem-solving accuracy. In line with this narrative, each facial posture task was followed by a brief single-digit math problem and a set of Likert-scale self-report questions asking about perceived task difficulty. This procedure was intended to minimize demand characteristics and reduce the likelihood of participants inferring the true purpose of the study.

The dependent variable was self-reported happiness, computed as a composite of four items: happiness, enjoyment, satisfaction, and liking, each rated on a 7-point Likert scale. Following the task phase, participants completed a series of post-experiment measures, including an awareness probe, quality indicators (e.g., compliance, similarity, and genuineness ratings), and an adapted DEELS subscale measuring the frequency of smiling without feeling happy. An additional set of exploratory questions assessed participants' judgment strategies during emotional self-reporting.

Materials

The visual stimuli for the primary tasks consisted of a 2×2 image collage of smiling faces, originally constructed by Coles et al. (2022) using images from the Extended Cohn–

Kanade Dataset (CK+; Lucey et al., 2010). An additional image of a woman smiling from the same dataset was used for the post-task similarity rating, along with a provided mirror.

To measure emotional response, participants rated their experience after each posture task using four Likert-scale items, happiness, enjoyment, satisfaction, and liking, rated from 1 (not at all) to 7 (an extreme amount), drawn from the happiness subscale of the Discrete Emotions Questionnaire (DEQ; Harmon-Jones et al., 2016). Additional post-experiment measures included an adapted subscale from DEELS, assessing the frequency of smiling without feeling happy (Appendix A). The items were adapted from their original workplace context to refer more broadly to social interactions. The original faking positive emotion subscale demonstrated high internal consistency (Cronbach's $\alpha = .87$). A set of semi-structured exploratory questions was also included to probe participants' judgment strategies and emotional appraisal processes during self-report (Appendix B).

All data were collected using the same personal computer. The criterion for the location was that participants would not be observed and that the room would be free from apparent distractions. This requirement led me to opt out of the distraction check present in the original study.

Procedure

Participants were seated at the designated computer workstation and were required to remain in a distraction-free environment, unobserved by others, throughout the experiment. They were informed that all instructions would be displayed on the screen in English; however, responses could be submitted in either English or Swedish. Before exiting the room, the investigator showed the participant a mirror and explained that instructions for its use would appear at the relevant stages of the experiment. The initial screen displayed a combined information sheet and consent form, which stated the study's purpose: to investigate the effects of movements and cognitive distractors on the accuracy and speed of solving mathematical problems.

After confirming that they had read the page and agreed to participate, participants undertook a series of four movement tasks, each paired with a mathematical problem and three sets of Likert-scale questions. The first task required participants to place their left hand behind their head and blink their eyes once per second for 5 seconds, while the fourth task involved tapping their left leg with their right-hand index finger at the same rate. The second and third tasks, which were of primary interest, were presented in randomized order. During these tasks, participants were shown an image of four smiling people. For the neutral task, participants were instructed to maintain a blank facial expression for 5 seconds regardless of the actors' expressions. For the happy task, they were instructed to mimic the actors' facial expressions for the same duration.

Participants were given time to practice all tasks and, upon confirming their readiness, proceeded to perform the tasks as a countdown was displayed on the screen. Following each task, they solved a single-digit addition or subtraction problem. They were then asked to rate the extent to which they experienced a range of emotions while performing the physical movement task. Ratings were made on a seven-point Likert scale and included happiness, enjoyment, liking, satisfaction, anger, nervousness, tiredness, confusion, and worry. All nine emotions were presented on the same page, though their order was randomized for each trial. The response options on the Likert scale ranged from "not at all" to "an extreme amount." Participants were then asked to evaluate the difficulty of both the movement task and the math problem, with response options ranging from "extremely easy" to "extremely difficult." On the second and third tasks, an additional question was included, instructing participants to select "extremely difficult." This question served as an attention check, and a correct response was required for inclusion in the quantitative analysis. After completing the emotional and difficulty ratings, participants answered a final question assessing how much they liked the physical movement task.

Next, participants were sequentially asked three questions designed to assess their awareness of the study's true purpose. The questions were: "What do you think the purpose

of the study was? Please write 1–2 sentences.” “In addition to our interest in how body movements and cognitive distractors influence mathematical speed and accuracy, do you think there is anything else we may have been interested in?” “We were actually interested in one other thing besides how body movements and cognitive distractors influence mathematical speed and accuracy. Please describe 1–2 ideas for what this might be.” The responses to these questions were evaluated by two independent coders, who each rated the participants' awareness on a seven-point Likert scale. This assessment was used as a criterion for inclusion in the quantitative analysis, where only participants who were rated as completely unaware by both coders were included.

To conclude the replication portion of the experiment, participants provided their age and gender, followed by three questions, each answered on a seven-point Likert scale, used as quality indicators for the data. These three items corresponded to compliance, genuineness, and similarity, respectively. The first question asked participants to rate the degree to which they followed the instructions during the mimicry task. A rating of “not at all” excluded the participant from the quantitative analysis. For the second question, participants were instructed to repeat the mimicry task and then rate the extent to which they felt they were expressing happiness. The third question asked participants to repeat the physical task while looking in the mirror and rate the extent to which their facial expression matched that of the smiling individual shown in the reference image. Participants who selected “not at all” for this question were excluded from the quantitative analysis.

Next, participants answered the three DEELS questions (see Appendix A), followed by nine questions for the qualitative analysis (see Appendix B). Finally, a debrief screen was presented, outlining the true purpose of the study, requesting that participants refrain from discussing the experiment during the two-week data collection period, and informing them that they could now ask the investigator any questions.

Analyses

Four self-report items, happiness, enjoyment, satisfaction, and liking, were averaged to create a composite happiness score for each condition. Awareness-related responses were rated independently by two coders who received instructions identical to those used in the original study. Only participants whom both coders had rated as unaware of the study's true purpose were included in all quantitative analyses, except the awareness regression. Exploratory qualitative responses were coded by the author and used to inform interpretation, but not for hypothesis testing.

All statistical analyses were conducted using R version 4.3.2 (R Core Team, 2023). Analyses for hypotheses H1 and H3–H7 were pre-specified and executed without post hoc model modifications, using a pre-coded R Markdown script. The specific models associated with each hypothesis are detailed in the Results section. Hypothesis H2 was also theory-driven, predicting fewer extreme values compared to the original study, but the analytic approach was not pre-specified. The method used (percentile comparison) was selected after inspecting the data and is therefore considered exploratory.

Although this study evaluates seven hypotheses, only one (H3) is newly introduced. The remaining hypotheses (H1, H4–H7) were directly replicated from Coles et al. (2022), and H2 was explicitly exploratory. This design allows for a focused test of a single novel theoretical contribution while maximizing comparability with prior findings.

Results

Twenty-four participants who met all predefined inclusion criteria were included in the analyses. The mean happiness rating in the present study was 2.63 ($SD = 1.52$) in the happy condition and 2.20 ($SD = 1.09$) in the blank condition. These means were further divided by the order in which the main tasks were presented, as shown in Table 1.

Table 1*Means of Happiness Ratings Based on Task Order*

Task Order	1 st task	2 nd task	3 rd task	4 th task
BlankFirst	2.69 (1.36)	1.73 (0.68)	2.38 (1.52)	2.21 (1.49)
MimicFirst	2.77 (1.25)	2.88 (1.55)	2.67 (1.24)	3.06 (1.35)

Note. In the 'Blank first' category ($n = 12$), the 2nd task is the blank expression, and the 3rd task is the mimicry condition. For 'Mimic first' ($n = 12$), the 2nd task is the mimicry condition, and the 3rd task is the blank expression. Numbers in parentheses represent standard deviations.

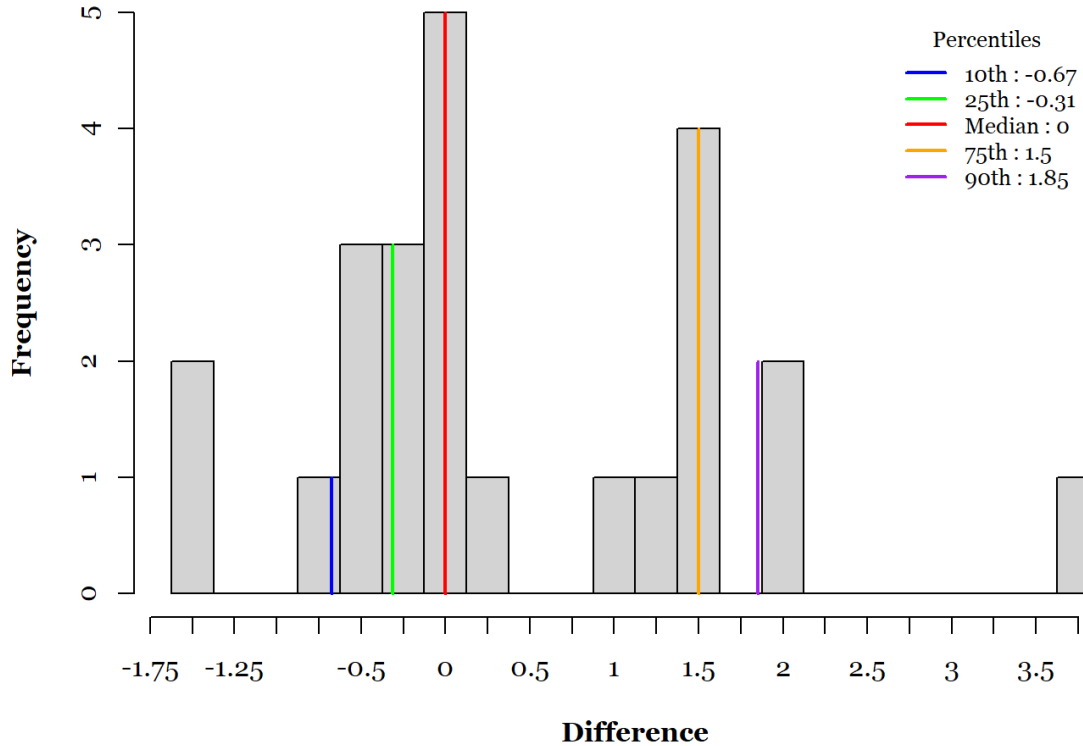
To explore potential order effects, a mixed-design ANOVA was conducted with task (four levels: filler1, blank, mimic, filler4) as a within-subjects factor and group (BlankFirst vs. MimicFirst) as a between-subjects factor. The analysis revealed no significant main effect of group, $F(1, 22) = 1.62, p = .217$, and no significant group \times task interaction, $F(3, 66) = 1.48, p = .227$. The main effect of task approached significance, $F(3, 66) = 2.20, p = .096$.

Although the interaction between task and group was not statistically significant, the means presented in Table 1 suggested potential group differences, particularly in the blank condition. Therefore, an exploratory set of pairwise comparisons was conducted using estimated marginal means with Tukey-adjusted p-values to examine group differences within each task. These comparisons tested for group differences within each task. A significant group difference was observed in the blank condition, $p = .032$. No other comparisons reached statistical significance.

The primary interest of the study, the distribution of difference magnitude in happiness ratings between the blank expression and smiling task among the 24 included participants, is illustrated in Figure 1.

Figure 1

Histogram Showing the Difference in Happiness Rating: Smiling-Blank Expression



Note. $n = 24$. The data include all participants in the present study who met all inclusion criteria. For the same graph using data from the original study, see Appendix C.

Inferential Analyses

To test H1, a Wilcoxon signed-rank test was conducted to compare happiness ratings between the happy and blank facial expression conditions. The median happiness rating was higher in the happy condition ($Mdn = 2.38$, $IQR = 1.81$) than in the blank condition ($Mdn = 1.75$, $IQR = 1.63$). However, this difference was not statistically significant, $W = 129.5$, $p = .085$, $r = .28$.

To test H2, an exploratory binomial analysis was conducted to assess whether the replication sample contained fewer tail-end scores than observed in the original study. Thresholds of -0.5 and 2.5 represented the 10th and 90th percentiles of the original study's distribution of happiness difference scores. In the current sample, 4 of 24 participants

(16.7%) fell outside this range. In a sample of this size, a statistically significant reduction would require observing 1 or fewer such cases. As the observed count exceeded this threshold, the result was not statistically significant.

To test H3, a multiple linear regression was conducted with fake smiling and genuine smiling as predictors of happiness difference scores. The model was not statistically significant, $F(2, 21) = 1.081$, $p = .358$, and explained approximately 9.3% of the variance ($R^2 = .093$). Neither fake smiling ($p = .571$) nor genuine smiling ($p = .178$) significantly predicted happiness differences.

To test H4, a simple linear regression was conducted with awareness as the predictor of happiness difference scores. The model was not statistically significant, $F(1, 63) = 1.53$, $p = .221$, and explained only 2.4% of the variance ($R^2 = .024$).

To test H5–H7, a multiple linear regression was conducted with compliance, similarity, and genuineness as predictors of happiness difference scores. The model was not statistically significant, $F(3, 20) = 1.92$, $p = .159$, although it explained approximately 22% of the variance ($R^2 = .224$). Of the three predictors, only genuineness significantly predicted happiness differences; compliance and similarity were not significant. Full regression results are presented in Table 2.

Table 2

Regression results for H3–H7 predictors of happiness difference

H #	Predictor	B	95% CI	df	t	p
H3	Fake smiling	-.123	[-.565, .320]	21	-.576	.571
–	Genuine smiling	.434	[-.213, 1.081]	21	1.394	.178
H4	Hypothesis awareness	.099	[-.061, .258]	63	1.237	.221
H5	Compliance	.155	[-.431, .741]	20	0.552	.587
H6	Similarity	.007	[-.333, .320]	20	-0.043	.966
H7	Genuineness	.355	[.026, .685]	20	2.250	.036

Note. All coefficients are unstandardized (B). Degrees of freedom refer to the residual df used for the t-tests in each model. Predictors for H3 and H4 were tested in separate regression models, while predictors for H5–H7 were tested together in a single model.

Exploratory Observations from Open-Ended Responses

To complement the quantitative results, responses to post-experiment open-ended questions were analyzed for themes that could shed light on individual rating patterns. From the current study's complete data set of 65 participants, five had a happiness rating that was more than 2.75 higher for the smiling condition compared to the blank facial expression condition. Of those participants, one stated that he was unsure if he should answer how he felt in general while performing the tasks or how the tasks had made him feel for the moment, and therefore, he felt inconsistent in his answers. Another participant revealed that it had felt very unnatural not to smile while seeing others smile. One stated that he had been reminiscing about how he had felt happy in the past days while mimicking the smile, and one responded that his emotions were not as per the rating.

Among the rest of the participants, another four reported having reflected on past emotions or events while assessing their emotion, but unlike the previously mentioned instance, these four did not explicitly connect the reflection to the smiling condition. Five participants had a happiness rating that was more than 1.0 lower in the smiling condition. One participant explained that for him, it was 'funny to stay calm while the others are smiling or laughing.' Another expressed that she thought she had rated how happy she was to do the movement task, that she did not feel that her happiness was different, and that her ratings did not represent her actual emotions well.

Discussion

The main purpose of the present study was to explore potential factors that may influence the distribution of individual responses observed in FFH designs, such as the one utilized in Coles et al. (2022). Understanding the sources of this distributional pattern is essential if the FFH is ever to be developed as a reliable intervention (Cronbach, 1957). Factors investigated included measurement-related methodological factors that may introduce systematic errors between conditions, such as potential order effects, differences in appraisal strategies, task misunderstandings, and contextual cues (Shadish, Cook, &

Campbell, 2002), as well as individual-difference factors, such as reported history of frequent emotional labor, which was examined as a potential moderator of the facial feedback effect. To enable this investigation, the study employed a partial replication of the facial mimicry without positive stimuli condition from Coles et al. (2022). The study was exploratory in nature and included hypothesis-generating components, such as participants' post-task reflections and open-ended questions about rating strategies, to identify factors that could inform improvements in the design and interpretation of facial feedback studies.

Exploratory testing indicated a significant order effect, whereas self-reported history of frequent emotional labor did not moderate the facial feedback effect. Conducting the study in person did not clearly reduce the occurrence of high within-participant variation compared to the original online study. The study's exclusion rate for the attention checks was higher than expected, which reduced statistical power. This reduction likely explains the results from the group-level estimate of the facial feedback effect, which was in the expected direction and close to reaching statistical significance. The findings showed concerning differences in task interpretation and appraisal strategies, suggesting factors that may warrant further investigation in future research on facial feedback effects.

Order Effects

Counterbalancing is frequently cited as a method to mitigate order effects, but it is important to understand that counterbalancing does not necessarily safeguard against such effects (Keren, 1992). Importantly, counterbalancing offers a valuable opportunity to systematically analyze data for the presence of order effects. An exploratory analysis revealed a statistically significant difference in the blank condition. A visual inspection of the data (see Table 1) showed a pattern reminiscent of an anchoring effect with minor adjustments starting at the third task. This type of anchoring-and-adjustment heuristic has been found specifically to operate on self-generated anchor values and generate insufficient adjustments (Epley & Gilovich, 2006). Why this hypothetical anchoring seemed to arise only after the second task, and not the first, is still unexplained.

Open-Ended Reports of Reflections

A core component in facial feedback studies is to obtain accurate self-reports of participants' emotional states. Teoh et al. (2023) frame such reports as dynamic affective decisions rather than as simple readouts of internal states. This decision-making process uses several sources of input and is vulnerable to variability from noise and bias, yet some of this variability may be shaped by identifiable moderators. Robinson and Clore (2002a) propose that emotional self-reports can be influenced by different sources of knowledge, including experiential, episodic, situation-specific, and semantic, and that the relative contribution of each depends on which source is most accessible at the time of judgment.

This distinction is directly relevant to the present study, where five participants reported having reflected on past emotions or events while assessing their emotion. From the perspective of Robinson and Clore's (2002a) accessibility model, this means that the most accessible source at the time of judgment was likely not the intended direct experiential access, but recalled episodes, leading to ratings that do not reflect the affective state evoked by the facial posture itself. This concern is not new; Schnall and Laird (2003), in their "Keep Smiling" study, explicitly instructed participants not to think of happy memories but to act happiness only with their face and body, in anticipation of precisely this type of confound. This issue is potentially more serious when past events are vague or not recent, because judgments may instead be influenced by semantic knowledge and produced through a qualitatively different process (Robinson & Clore, 2002b). This poses a serious validity concern, because if participants rely on different judgment processes, then ratings cannot be assumed to reflect the same underlying construct or causal pathway, but may instead represent a mixture of expression-driven affect and memory-based judgment.

Although five participants reported reflecting on past events, the true prevalence of this tendency is unknown, and it is likewise uncertain how many may have done so selectively by condition, a bias explicitly reported by one participant. This type of systematic bias is especially concerning, because it not only means that participants may draw on different

sources of knowledge across conditions, but also that judgments in different conditions may be generated through different processes (Robinson & Clore, 2002a, 2002b). If reflections influence ratings, investigating the mechanisms involved may be best pursued by considering multiple theoretical perspectives.

It is possible that such reflections act as moderators of felt emotion, either by influencing appraisal processes or even by functioning as part of the appraisal processes that contribute to emotional experience (Moors, 2013). Another possibility is that they emerge primarily when participants are asked to provide self-reports, in which case they reflect self-report reactivity rather than changes in the emotion itself (Kassam & Mendes, 2013). From this perspective, reflections moderate the rating process rather than the felt experience, consistent with the view that self-reports are dynamic affective decisions shaped by multiple inputs (Teoh et al., 2023). Reflections on the past threaten construct validity to the extent that participants' ratings reflect memory-based judgment. However, if reflections influence felt affect and, consequently, the ratings, the expression-driven change in felt affect we hope to measure would instead reflect a memory-driven change in felt affect triggered by the expression, which would threaten internal validity.

Cognitive Strategies in Emotion Ratings

Open-ended responses indicated that more than a quarter of participants judged at least some of their later ratings in relation to their previous ratings instead of or in combination with making a new affective assessment. Comparable distinctions have been made by Gavanski (1986), who showed that cognitive evaluations of funniness and affective assessments of amusement can diverge. Strack et al. (1988) applied this distinction in a facial feedback paradigm, finding effects only on affective ratings of amusement and not on cognitive ratings of funniness, even though both were made for the same cartoons. If some participants in the present study based their ratings on cognitive comparisons, their reports may have reflected a different construct from that produced by affective assessments, thereby threatening construct validity.

Adapted DEELS as Moderator

Given the study's limited power to detect a statistically significant moderation by self-reported frequency of emotional labor, at most an indication of moderation was anticipated. Since no such indication appeared, one possibility is that the proposed mechanism does not exist. Another possibility is that the mechanism exists but that contextual factors in the present study may have attenuated the effect (Keller et al., 2023). A third possibility is that the relevant conditioning process is most plastic in childhood and becomes less modifiable with age (Braun & Geiselhart, 1959). If so, self-reported history of emotional labor in adulthood would likely be a weak moderator, and developmental factors such as attachment patterns could be more informative targets for future investigation (Ainsworth et al., 1978). Lastly, the concerns about construct and internal validity discussed earlier raise the possibility that the present study may not have measured facial feedback as intended.

Strengths

Two important additions to the original design enabled exploratory analyses of potential moderators and methodological artifacts. First, the study recorded task order, enabling a pre-planned exploratory analysis of order effects to evaluate whether sequencing influenced ratings rather than presuming that counterbalancing eliminated such influences. Second, a semi-structured post-experiment questionnaire with open-ended prompts about how participants formed their ratings provided additional data on similarities and differences in rating strategies and other influences on their emotion ratings. These additions are important tools for evaluating the study's validity.

Limitations

The present study was underpowered for most planned analyses, compounded by a higher-than-expected exclusion rate due to failed attention checks, reducing both statistical power and precision. Most analyses were exploratory and should be treated as potential guides for future designs rather than as results on their own. The post-experiment responses

of participants who reported having made later emotion ratings in relation to previous ratings were partially interpreted rather than pre-coded, which makes both the prevalence of that tendency and the extent of its influence very uncertain. Narrower and more explicit questions could have yielded more reliable results, but would also have reduced the exploratory scope.

The qualitative post-experiment prompts were explicitly exploratory and were not coded with a formal scheme. Therefore, any prevalence indications derived from open-ended responses must be treated as hypothesis-generating rather than definitive. Although the study recorded task order and examined it, these analyses were exploratory, so the study cannot adjudicate whether sequence effects are absent or present. Unlike the image viewed during the main tasks, the image of a girl smiling shown while participants assessed the similarity rating lacked visible teeth, which may have affected similarity judgments and exclusions. Also, some participants reported that the smiles shown during the main tasks seemed fake, which may have influenced both how they smiled and how they rated genuineness and similarity.

Hiding the true purpose among other irrelevant tasks is a good way to minimize potential demand characteristics. The only time in an experiment when demand characteristics can influence the results is up until the last measured variable of interest. Therefore, the fourth task could not possibly be used to mitigate such effects. However, a participant who suspected the purpose of the study and was influenced by demand characteristics at the rating of the third task may be confused by the irrelevance of the fourth task and therefore not report the once-suspected purpose. In other words, the fourth task can only confuse participants ahead of the inclusion criteria report, thus only serving as a way to include participants who may already have been influenced by demand characteristics.

Since the participants were recruited through in-person solicitation, there is also a possibility of sampling bias. There was a clear preference for recruiting participants who were

not in a group. This preference aimed to minimize the potential influence of stress from others waiting, which could affect participants' responses.

Implications and Future Research

Five participants reported having reflected on past events while rating their emotions, and the prevalence and effects of these reflections should be assessed. This can be done by including closed-ended questions asking whether and in which conditions reflection occurred, as well as questions to categorize the type of reflection in each condition. Pre-registered analyses should test whether the mean within-participant difference in ratings differs between participants who reflected and those who did not, and whether reflection type moderates that difference.

The construct validity concerns regarding the comparative strategies, conceptualized as a form of panel conditioning (Bergmann & Barth, 2017), in which participants judged their later ratings in relation to earlier ones require further investigation. The Visual Analogue Scale (VAS), which may be advantageous over Likert-type scales for affective assessment (Haslbeck et al., 2025; Sung & Wu, 2018), has been proposed as a means to mitigate such comparative strategies (Monk, 1989). Future work should test whether a VAS implementation designed to minimize comparative strategies changes the distribution of within-participant differences in ratings, and whether any such change is stronger among participants who report comparing later with earlier ratings.

Lastly, because the analysis that detected the order effect was exploratory, future work should record task order and pre-register a confirmatory test of order effects.

Conclusion

The present exploratory study examined methodological and individual-difference factors that may shape the distribution of individual responses in FFH designs. The findings do not indicate a significant difference between online and in-person testing. While the facial feedback effect was in the same direction as the original study, it did not quite reach

statistical significance. All quality indicators showed a pattern similar to that in the original study, but only one, genuineness, reached statistical significance. Self-reported frequency of emotional labor was not found to moderate the facial feedback effect.

A substantial portion of the sample reported making later emotion ratings in relation to their previous ratings, and five participants reported thinking about past events while rating their emotions. These reports raise serious concerns regarding the study's internal validity. Suggested priorities for future work include closed-ended questions to assess the prevalence and influence of reflection, evaluation of a VAS implementation designed to minimize comparison strategies, and a pre-registered confirmatory test of order effects.

Final word count: 7500 words.

References:

- Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum.
- Bergmann, M., & Barth, A. (2017). What was I thinking? A theoretical framework for analysing panel conditioning in attitudes and (response) behaviour. *International Journal of Social Research Methodology*, 21(3), 333–345.
<https://doi.org/10.1080/13645579.2017.1399622>
- Borg, C., de Jong, P. J., & Leus, I. (2016). Is disgust sensitive to classical conditioning as indexed by facial electromyography and behavioural responses? *Cognition and Emotion*, 30(4), 687–698. <https://doi.org/10.1080/02699931.2015.1022512>
- Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry*, 52(10), 976–986. [https://doi.org/10.1016/S0006-3223\(02\)01546-9](https://doi.org/10.1016/S0006-3223(02)01546-9)
- Braun, H. W., & Geiselman, R. (1959). Age differences in the acquisition and extinction of the conditioned eyelid response. *Journal of Experimental Psychology*, 57(6), 386–388.
<https://doi.org/10.1037/h0039206>
- Buck, R. (1980). Nonverbal behavior and the theory of emotion: The facial feedback hypothesis. *Journal of Personality and Social Psychology*, 38(5), 811–824.
<https://doi.org/10.1037/0022-3514.38.5.811>
- Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L. G., Willis, M. L., Foroni, F., Reggev, N., Mokady, A., Forscher, P. S., Hunter, J. F., Kaminski, G., Yüvrük, E., Kapucu, A., Nagy, T., Hajdu, N., Tejada, J., ... Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature Human Behaviour*, 6(12), 1731–1742.
<https://doi.org/10.1038/s41562-022-01458-9>
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12(11), 671–684. <https://doi.org/10.1037/h0043943>
- Darwin, C. (1872). *The expression of the emotions in man and animals*. John Murray.
- De Houwer, J. (2020). Conditioning is more than association formation: On the different ways in which conditioning research is valuable for clinical psychology. *Collabra: Psychology*, 6(1), Article 2. <https://doi.org/10.1525/collabra.239>
- Delle Fave, A., Brdar, I., Freire, T., Vella-Brodrick, D., & Wissing, M. P. (2011). The eudaimonic and hedonic components of happiness: Qualitative and quantitative

- findings. *Social Indicators Research*, 100(2), 185–207.
<https://doi.org/10.1007/s11205-010-9632-5>
- Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, 55(1), 34–43. <https://doi.org/10.1037/0003-066X.55.1.34>
- Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58(3), 403–417. <https://doi.org/10.1093/biomet/58.3.403>
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3–4), 169–200. <https://doi.org/10.1080/02699939208411068>
- Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, 17(4), 311–318. <https://doi.org/10.1111/j.1467-9280.2006.01704.x>
- Folk, D., & Dunn, E. W. (2024). How can people become happier? A systematic review of preregistered experiments. *Annual Review of Psychology*, 75, 467–493. <https://doi.org/10.1146/annurev-psych-022423-030818>
- Gavanski, I. (1986). Differential sensitivity of humor ratings and mirth responses to cognitive and affective components of the humor response. *Journal of Personality and Social Psychology*, 51(1), 209–214. <https://doi.org/10.1037/0022-3514.51.1.209>
- Glomb, T. M., & Tews, M. J. (2004). Emotional labor: A conceptualization and scale development. *Journal of Vocational Behavior*, 64(1), 1–23. [https://doi.org/10.1016/S0001-8791\(03\)00038-1](https://doi.org/10.1016/S0001-8791(03)00038-1)
- Harmon-Jones, C., Bastian, B., & Harmon-Jones, E. (2016). The Discrete Emotions Questionnaire: A new tool for measuring state self-reported emotions. *PLOS ONE*, 11(8), e0159915. <https://doi.org/10.1371/journal.pone.0159915>
- Harris, J. A., Jones, M. L., Bailey, G. K., & Westbrook, R. F. (2000). Contextual control over conditioned responding in an extinction paradigm. *Journal of Experimental Psychology: Animal Behavior Processes*, 26(2), 174–185. <https://doi.org/10.1037/0097-7403.26.2.174>
- Haslbeck, J. M. B., Jover Martínez, A., Roefs, A. J., Fried, E. I., Lemmens, L. H. J. M., Groot, E., & Edelsbrunner, P. A. (2025). Comparing Likert and visual analogue scales in ecological momentary assessment. *Behavior Research Methods*, 57, Article 217. <https://doi.org/10.3758/s13428-025-02706-2>

- Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. <https://doi.org/10.1037/a0018916>
- James, W. (1884). What is an emotion? *Mind*, *9*(34), 188–205. <https://www.jstor.org/stable/2246769>
- Kassam, K. S., & Mendes, W. B. (2013). The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. *PLOS ONE*, *8*(6), e64959. <https://doi.org/10.1371/journal.pone.0064959>
- Keller, N. E., Cooper, S. E., McClay, M., & Dunsmoor, J. E. (2023). Counterconditioning reduces contextual renewal in a novel context but not in the acquisition context. *Neurobiology of Learning and Memory*, *201*, Article 107749. <https://doi.org/10.1016/j.nlm.2023.107749>
- Keren, G. (1992). Between- or within-subjects design: A methodological dilemma. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 257–272). Lawrence Erlbaum.
- Kleck, R. E., Vaughan, R. C., Cartwright-Smith, J., Vaughan, K. B., Colby, P. M., & Lanzetta, J. T. (1976). Effects of being observed on expressive, subjective, and physiological responses to painful stimuli. *Journal of Personality and Social Psychology*, *34*(6), 1211–1218. <https://doi.org/10.1037/0022-3514.34.6.1211>
- Kringelbach, M. L., & Berridge, K. C. (2010). The neuroscience of happiness and pleasure. *Social Research: An International Quarterly*, *77*(2), 659–678.
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops* (pp. 94–101). IEEE. <https://doi.org/10.1109/CVPRW.2010.5543262>
- Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175–190. <https://doi.org/10.1037/1528-3542.5.2.175>
- Monk, T. H. (1989). A visual analogue scale technique to measure global vigor and affect. *Psychiatry Research*, *27*(1), 89–99. [https://doi.org/10.1016/0165-1781\(89\)90013-9](https://doi.org/10.1016/0165-1781(89)90013-9)
- Moors, A. (2013). On the causal role of appraisal in emotion. *Emotion Review*, *5*(2), 132–140. <https://doi.org/10.1177/1754073912463601>

- Nitschinsk, L., Tobin, S. J., & Vanman, E. J. (2022). The disinhibiting effects of anonymity increase online trolling. *Cyberpsychology, Behavior, and Social Networking*, 25(6), 377–383. <https://doi.org/10.1089/cyber.2022.0005>
- 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 and Social Psychology*, 114(5), 657–664. <https://doi.org/10.1037/pspa0000121>
- R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Robinson, M. D., & Clore, G. L. (2002a). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6), 934–960. <https://doi.org/10.1037/0033-2909.128.6.934>
- Robinson, M. D., & Clore, G. L. (2002b). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology*, 83(1), 198–215. <https://doi.org/10.1037/0022-3514.83.1.198>
- Sander, D. (2013). Models of emotion: The affective neuroscience approach. In J. L. Armony & P. Vuilleumier (Eds.), *The Cambridge handbook of human affective neuroscience* (pp. 5–45). Cambridge University Press. <https://doi.org/10.1017/CBO9780511843716.003>
- Schnall, S., & Laird, J. D. (2003). Keep smiling: Enduring effects of facial expressions and postures on emotional experience. *Cognition and Emotion*, 17(5), 787–797. <https://doi.org/10.1080/02699930302286>
- Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. Free Press.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, 11(6), 929–930. <https://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(5), 768–777. <https://doi.org/10.1037/0022-3514.54.5.768>

- Sung, Y.-T., & Wu, J.-S. (2018). The Visual Analogue Scale for Rating, Ranking, and Paired-Comparison (VAS-RRP): A new technique for psychological measurement. *Behavior Research Methods*, *50*(4), 1694–1715. <https://doi.org/10.3758/s13428-018-1041-8>
- Teoh, Y. Y., Cunningham, W. A., & Hutcherson, C. A. (2023). Framing subjective emotion reports as dynamic affective decisions. *Affective Science*, *4*, 522–528. <https://doi.org/10.1007/s42761-023-00197-y>
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., ... Zwaan, R. A. (2016). Registered replication report: Strack, Martin, and Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. <https://doi.org/10.1177/1745691616674458>
- Wänke, M., & Schwarz, N. (1997). Reducing question order effects: The operation of buffer items. In L. E. Lyberg, P. P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 115–140). Wiley.

Appendix A

Semi-Structured Post-Experiment Questionnaire - DEELS Component

<p>Almost done - just a few more questions Please take your time with these, as they are the core of our study</p>					
<p>We would like to know about the happiness emotions you express to others, such as customers, clients, coworkers, supervisors, etc., and emotions that you feel but do not express. That is, we are interested in the conditions and frequency in which you express happiness, specifically through your facial expressions. The following sections may seem somewhat similar, so please read the instructions carefully and consider your experiences over the past six months.</p>					
<p>How often do you genuinely express happiness when you feel that way?</p>					
I genuinely express happiness many times a day	I genuinely express happiness a few times a day	I genuinely express happiness a few times a week	I genuinely express happiness a few times a month	I never genuinely express happiness	
<p>How often do you express feelings of happiness when you really do not feel that way?</p>					
I express happiness many times a day when I do not feel it	I express happiness a few times a day when I do not feel it	I express happiness a few times a week when I do not feel it	I express happiness a few times a month when I do not feel it	I never express happiness when I do not feel it	
<p>How often do you keep feelings of happiness to yourself when you really feel that way?</p>					
I keep happiness to myself many times a day	I keep happiness to myself a few times a day	I keep happiness to myself a few times a week	I keep happiness to myself a few times a month	I never keep happiness to myself	I never feel happiness

Appendix B

Semi-Structured Post-Experiment Questionnaire - Emotion Rating Component

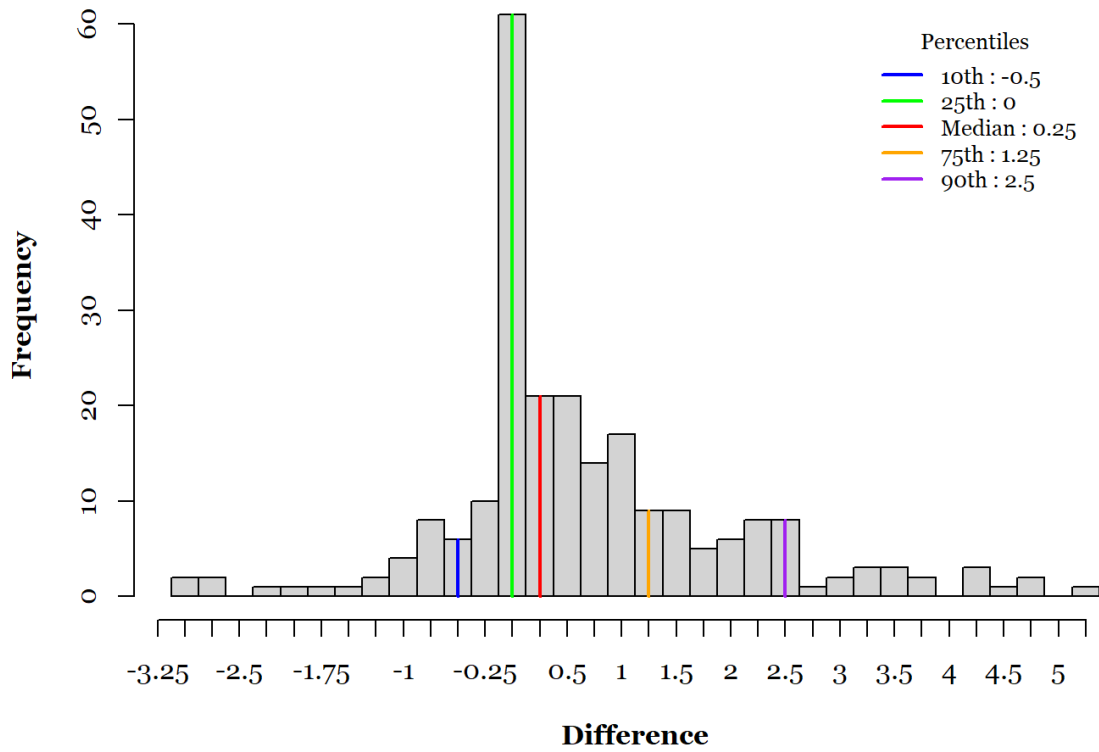
How difficult was it to accurately represent your emotions on the 7-point scales?	
How well do you think your ratings represented your actual emotions at those times?	
How would you describe your method or process for determining the ratings for your emotions, and did the method differ depending on the emotion?	
After the first emotion assessment, did you choose your emotion levels in the following emotion questions in relation to how you had previously answered, or did you make a completely new assessment of your emotions?	
If higher	The average rating across four categories (happiness, enjoyment, satisfaction, and liking) was (<i>Difference</i>) higher after the smiling posture compared to after the blank posture. Does that seem to represent your emotions accurately at those times, and to what do you attribute the difference?
If same	The average rating across four categories (happiness, enjoyment, satisfaction, and liking) was unchanged between the measurements after the smiling posture and after the blank posture. Did you feel that there was a difference, but too small for a whole step on the 7-point scale?
If lower	The average rating across four categories (happiness, enjoyment, satisfaction, and liking) was (<i>Difference</i>) lower after the smiling posture compared to after the blank posture. Does that seem to represent your emotions accurately at those times, and to what do you attribute the difference?
Do you feel that you experienced an emotional difference between the ratings that did not fit into any of the categories? If so, how would you describe it?	
If you try to generate a genuine smile now, do you feel any change in emotion? If so, how would you describe it?	
Estimate the range of times you smile genuinely, on average, per week. (e.g. 4-9)	
<p style="text-align: center;">Final Questions:</p> How would you rate your mood at the start of the experiment? If there has been a change, how would you rate your mood right now? (Two 9-point Likert scales with five labels `Very Negative` – `Somewhat Negative` – `Neutral` – `Somewhat Positive` – `Very Positive`)	

Appendix C

Histogram of Happiness Rating Differences: Blank Expression vs. Smiling

Figure 2

Happiness Rating Differences



Note. $N = 235$. The data include all participants from the original study who met all inclusion criteria, as defined by Coles et al. (2022). Difference = Smiling – Blank.