

Philadelphia's Got Talent! A novel, naturalistic paradigm for inducing socio-affective prediction errors

Linda J. Hoffman

Temple University

Nicole Serino

nicole.serino@temple.edu

Temple University

Steven Martinez

Temple University

Morgan Callinan

Temple University

Zack Mulron

Temple University

Chelsea Helion

Temple University

Vishnu P. Murty

University of Oregon

Ingrid R. Olson

Temple University

Research Article

Keywords: Prediction error, Embarrassment, Liking gap, Social norm violation

Posted Date: September 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10656829/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Philadelphia's Got Talent! A novel, naturalistic paradigm for inducing socio-affective prediction errors

Abstract

Prediction error — the discrepancy between what is expected and what actually occurs — is increasingly recognized as a core computational principle organizing perception, action, and affect. Embarrassment can be conceptualized as a socio-affective prediction error, arising when social outcomes violate expectations. Here, we introduce a task to systematically induce embarrassment and measure these prediction errors, called Philadelphia's Got Talent (PGT). In the first session, participants sang, danced, or told a silly story while being videotaped. A week later, they unexpectedly viewed footage of their own and another participant's performance while providing continuous ratings. Results show that participants underestimated how positively others would evaluate their performances — a "liking gap" — and exhibited socio-affective prediction errors specifically during norm-violating performances, but not during control conditions. Participants self-reported feeling embarrassed by their own performances but amused by other people's performance. Together, these findings support PGT as a paradigm for eliciting and quantifying socio-affective prediction errors, capturing individual differences in responses to social norm violations, and offering a computational framework for future research in social psychology and social neuroscience as well as clinical science.

Keywords: Prediction error, Embarrassment, Liking gap, Social norm violation

Introduction

Real-time social interaction requires one to quickly interpret ambiguous cues and misinterpretation can lead to accidental violations of social norms and expectations. Detecting and correcting these errors rapidly is essential for restoring social integrity, making predictive mechanisms particularly important to social interaction (Mayr, 1974; Brown & Brüne, 2012). Within a predictive processing framework, the brain uses prior experience to construct internal models—learned mental representations of regularities in the external world—that anticipate future states and likely outcomes, allowing new information to be interpreted through existing expectations (Friston et al., 2006; Bar, 2007; Ito, 2008). As new information arrives, predictions generated by these models are compared with incoming sensory input, and mismatches between predicted and actual outcomes produce prediction errors (PEs) that signal when the models should be updated to support rapid learning and adaptive behavior (Brown & Brüne, 2012; Ito, 2008). In the social domain, internal models incorporate culturally shaped social norms that guide predictions about how one’s behavior will be evaluated, such that PEs arise when outcomes violate expectations (Baron-Cohen et al., 1985; Allman et al., 2011; Brown & Brüne, 2012). Such errors can trigger self-conscious emotions, such as embarrassment, which motivates immediate behavioral correction to preserve social integrity (Joiner et al., 2017; Frith & Frith, 2012). We define this combination of violated social expectations and an accompanying affective response as a socio-affective PE. Embarrassment therefore provides a useful model for experimentally studying socio-affective PEs.

Existing paradigms have separately examined social norm violations and experimentally induced embarrassment, providing an important methodological foundation. The Faux Pas Task, one of the most widely used measures of social norm violation detection, asks participants to

identify social missteps presented in written vignettes, cartoons, videos, or imagined socially inappropriate scenarios (Baron-Cohen et al., 1999; Krach et al., 2011; Mayer et al., 2020; Melchers et al., 2015; Moll et al., 2007). An alternative approach to study embarrassment has been emotion-induction paradigms, which aim to elicit embarrassment directly. For example, in the karaoke task, participants are recorded singing and then unexpectedly shown the recording, a procedure that reliably induces embarrassment (Brown & Garland, 1971; Shearn et al., 1990; Strum et al., 2008). More recent fMRI studies have induced embarrassment by giving participants manipulated negative performance feedback and informing them that their scores would be shown to the group (Mueller-Pinzler et al., 2015).

Although these paradigms provide important foundations, they were not designed to capture the full socio-affective PE process. First, they rarely examine the same social situation from both an egocentric perspective—experienced by the participant—and a vicarious perspective—observed in another person—making it difficult to isolate agent-dependent mechanisms underlying self-conscious emotions. Second, they typically lack a trial-by-trial computational measure of embarrassment-related PE magnitude. Third, they provide limited ability to track the temporal dynamics of PE during the induction. Finally, they may not reliably induce embarrassment across most participants.

To address these gaps, we developed *Philadelphia's Got Talent (PGT)*, a novel paradigm designed to naturalistically induce embarrassment while capturing multiple components of the socio-affective PE process. PGT consists of two sessions and requires minimal equipment and personnel. In PGT, participants performed embarrassing tasks, such as singing and dancing in front of a camera. One week later, they unexpectedly viewed recorded clips of their own performances, as well as recordings of a paired participant, in the presence of a real or perceived

audience. To capture moment-to-moment PEs, participants continuously rated how each video compared to their expectations during viewing. After each video clip, participants answered questions that allowed direct comparison between predicted and actual responses. Here, we report two studies: Study 1 provided a proof of concept that the PGT elicited measurable socio-affective PEs in a small sample, whereas Study 2 addressed methodological limitations of the initial iteration and assessed the robustness of these effects in a larger sample.

Our focal hypotheses concerned whether the PGT task elicited socio-affective PEs (see **Table 1**). Hypothesis 1 (H1) predicted a PE consistent with the liking gap, whereby people underestimate how positively others evaluate them, such that actual partner ratings would exceed predicted ratings (Boothby et al., 2018). The signed discrepancy between predicted and actual partner ratings indexed the PE's magnitude and direction. Hypothesis 2 (H2) predicted that participants would rate their performance more positively immediately after performing than after watching themselves on video, because self-observation was expected to reveal a discrepancy between how the performance felt and how it appeared. Hypothesis 3 (H3) predicted that participants would report more negative emotional valence when viewing themselves than when viewing another participant, because self-observation was expected to elicit greater self-conscious affect consistent with the negative valence of embarrassment (Keltner & Buswell, 1997; Miller, 1996; Singh & Bhushan, 2025).

Table 1. Summary of hypotheses, corresponding research questions, studies in which each question was tested, and whether each hypothesis was supported.

Hypothesis	Question	Study	Hypothesis Supported?
H1: Social PE "Liking Gap"	Q1. Will the partner rate the agent as better than expected (S2 - S1)?	1, 2	yes
	Q2. Will the partner rate the agent as better than expected (S2 - S2)?	1, 2	yes
	Q6. Will there be a liking-gap prediction error to passive control trials (S2 - S2)?	2	yes
H2: Egocentric PE "Doing vs Viewing"	Q3. Will the agent rate themselves as worse upon viewing compared to doing (S2 - S1)?	1, 2	no
	Q4. Will the predictions about partner's ratings change from doing to viewing (S2 - S1)?	1, 2	no
H3: Affective PE "Self vs Other"	Q5. Will participants feel more negatively about themselves than their partner (S2 - S2)?	2	yes
	Q7. Will participants feel more negatively about themselves than their partner during passive control trials (S2 - S2)?	2	no

We also examined subsidiary hypotheses. We hypothesized that greater creative freedom would amplify PE magnitude because less constrained performances introduce greater uncertainty, increasing the likelihood of a discrepancy between participants' expectations and how their performance appeared on video. We further predicted that PEs would be largest early in the experiment and diminish toward neutrality as participants habituated, consistent with PE-based learning accounts in which repeated exposure updates internal models and reduces uncertainty (Friston, 2010; Joiner et al., 2017). Finally, we included control conditions not intended to elicit PEs, allowing us to assess whether the observed effects were specific to the induction or generalized across viewing contexts.

Together, these studies represent an introduction and validation of PGT as a paradigm for eliciting and quantifying socio-affective PEs. In developing the task, we sought to balance

ecological validity with methodological flexibility, creating a paradigm that could be adapted for use across a range of research settings, including social psychological, clinical, and neuroimaging contexts. PGT may provide a useful framework for investigating features of socio-affective PEs.

Study 1

Study 1 Methods

Participants

Study 1 served as a proof-of-concept for the paradigm. We recruited 22 participants; six were lost due to technical and recording problems, and four were excluded because they or their dyad partner did not complete the second session. This resulted in a relatively small final sample of 12 participants forming 6 dyads. Participant demographics are reported in **Table S1** in Online Resource 1. Participants were recruited through the SONA systems for course credit, and through advertisements targeting residents of greater Philadelphia. Those who responded to study advertisements who did not receive SONA-based credit received monetary compensation for their participation. All participants provided consent in accordance with the Institutional Review Board. Participants were screened to ensure they had normal or corrected-to-normal vision and no diagnosis of neurological conditions, developmental delays, or cognitive disabilities. Finally, since completing the behavioral task in this study required cultural knowledge typically acquired in childhood, participants were also screened for attendance at kindergarten through middle school in the United States and for being native English speakers.

Event structure

PGT involves two sessions conducted one week apart. Before each task, participants provided consent and completed self-report questionnaires. In Session 1, participants completed a task called the Talent Show, which involved performing a series of tasks across different

performance modalities (e.g., song and dance) while recording video and audio. One week later, in Session 2, participants partook in the Surprise Self–Other Observation task (SSOO), during which they viewed videos of both themselves and another participant performing the tasks from Session 1.

Procedure

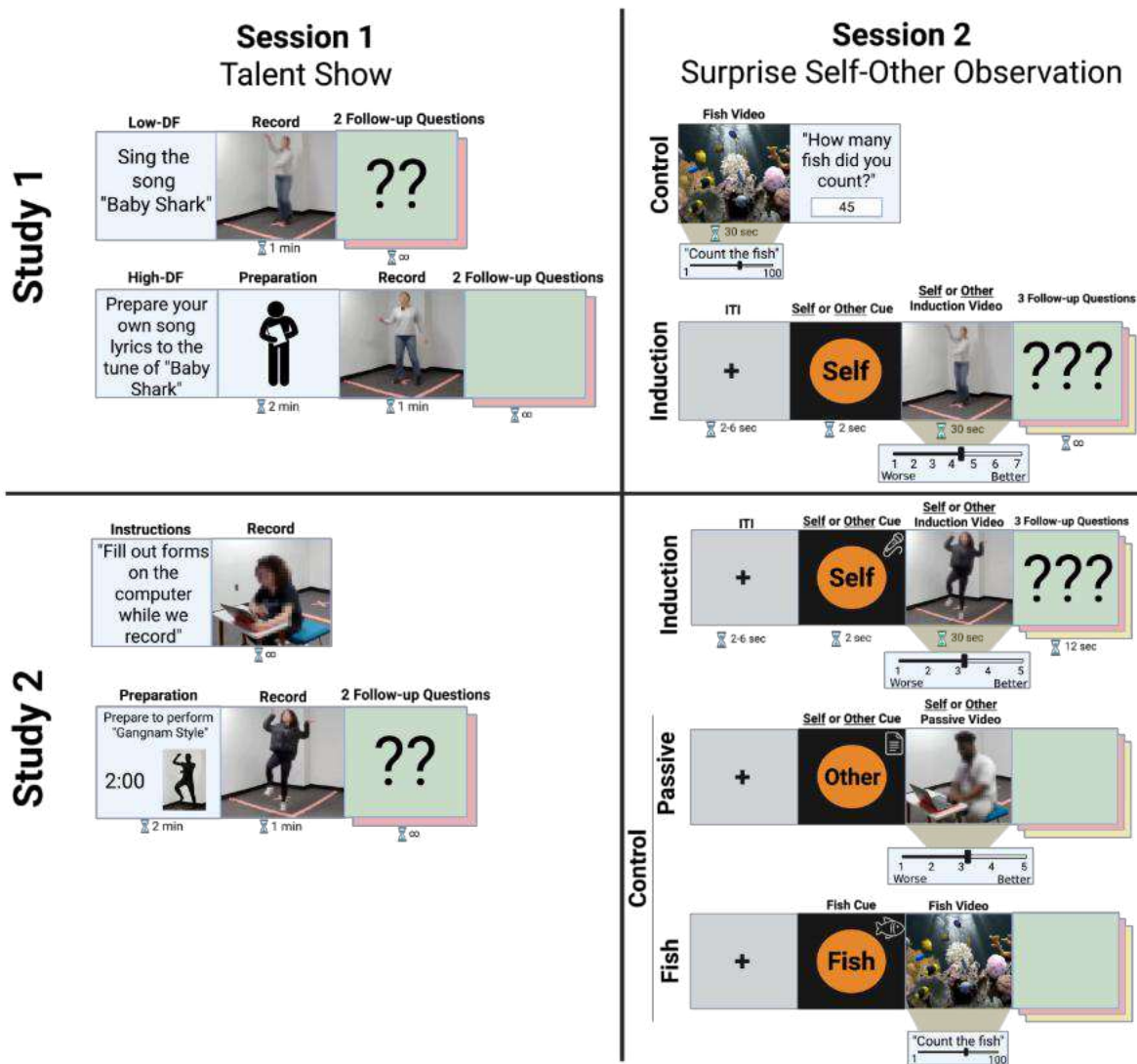


Fig. 1 Overview of experimental procedures for Study 1 (top row) and Study 2 (bottom row). In Session 1, participants in both studies completed the Talent Show, performing creative tasks while being recorded and answering follow-up questions after each trial. Study 1 included two

levels of creative freedom, whereas Study 2 replaced this manipulation with a passive recording of participants completing questionnaires. One week later, in Session 2, participants in both studies completed the Surprise Self–Other Observation task, viewing 30-second clips of their own and a partner's performances while providing continuous ratings and answering follow-up questions after each video. In Study 1, a fish control trial always preceded each performance trial; in Study 2, fish, passive, and performance trials were fully randomized. ITI = inter-trial interval.

Session 1: The Talent Show

Participants arrived at Session 1 of the PGT study under the impression that the study examined creativity. This cover story was relayed through recruitment information, consent materials, and verbal instructions provided at the start of the session. This deception was used to encourage authentic performances during the Talent Show and to prevent participants from suspecting that the task was designed to induce embarrassment. In Session 1, participants completed the Talent Show task (**Figure 1**) during which time they were videotaped performing twelve one-minute "creative" tasks across three performance modalities: storytelling, singing, and dancing. For each modality, participants were assigned varying degrees of creative freedom (DF) in how they performed the task. We hypothesized that greater creative freedom would amplify the embarrassment induction by manipulating subjects' agency over the performances. In the low-degrees-of-freedom trials (low-DF), participants followed a constrained prompt (e.g., "Sing Baby Shark for one minute in front of the camera"). In high-degrees-of-freedom trials (high-DF), participants had more time and freedom to generate their own variation of the same prompt (e.g., "Create your own song to the tune of Baby Shark"). The trial order was fixed across participants. Modalities were always presented in the following sequence: storytelling, singing, and dancing. Within each modality, two distinct tasks were performed, first under low-DF conditions and then under high-DF conditions. Each trial prompt was presented verbally by an experimenter from a printed face-sheet. By default, participants were assigned the first two prompts on the list unless they indicated they were unfamiliar with the task, in which case the experimenter reads through

the face-sheet until the participant confirmed a task they knew. While prompts were delivered verbally, the participant interacted with the task on a testing laptop projected onto a TV, which presented task instructions, timers, and questions for participants to engage with over the course of the task (described in detail below). Before starting the real task, a researcher walked through a practice sequence with the subject to ensure procedural understanding.

All trials required participants to perform in front of the camera for 1 minute, standing in a performance square taped on the floor in the corner of the testing room to ensure full-body visibility during recording. After delivering the prompt, the experimenter started a 70-second countdown (60 seconds for the performance, plus a 10-second buffer to exit the room), which was displayed on a TV screen adjacent to the camera. When the timer ended, it flashed red and emitted a tone to signal the end of the performance. Participants then rated their performance on a 7-point scale, with higher scores indicating more positive self-rating. Participants were also asked how much they would expect someone else to like their performance on a 7-point scale with higher scores indicating more optimistic expectations of social rating (see Appendix S3 for details in Online Resource 1).

High-DF trials included an additional component prior to the performance period. The high-DF trials consisted of an additional three-minute preparatory period that preceded the performance period. Participants were allotted this time during high-DF trials to generate their creative performances, and were given a clipboard, pen, and paper to assist them. A 3-minute timer was presented on the TV screen during this time. When time was up, the timer would flash red and make a sound. From there, a message appeared on the screen telling participants to get the experimenter for further instructions. The experimenter would then take the preparation materials (i.e., pen and paper) away from the subject so they could not rely on it during their performance.

Subsequently, subjects performed their high-DF performances in front of the camera as they did for the low-DF trials.

Session 2: Surprise Self-Other Observation task (SSOO)

Approximately one week later, naïve participants returned for the SSOO designed to induce embarrassment (**Figure 1**). Participants were intentionally not informed in advance about what Session 2 would involve, so that they could not prepare for or anticipate the self-observation component. To induce egocentric and vicarious embarrassment, participants were unexpectedly confronted with the video clips of their creative performances and those of another participant. Together, self and other performance clips will be referred to as *induction trials*. The self-condition for one subject was the other condition for their partner and vice versa. Participants were paired into dyads based on how close their sessions were in time. To amplify the embarrassment induction, an unfamiliar experimenter sat next to the participant throughout the task as an observer. The observer was instructed to remain silent and maintain a neutral facial expression throughout, so as not to indicate any reaction to the performances. Since embarrassment is a social phenomenon experienced in a real or imagined social context (Miller et al. 1992), the observer was used to elicit the social context necessary to promote the experience of embarrassment. An unfamiliar experimenter was used for this purpose to ensure that any impressions developed towards the experimenter in Session 1 did not influence the nature of the social context in Session 2. The structure and demands of the task are described in detail below.

The SSOO was presented on a TV screen linked to a response laptop, and the task was segmented into discrete blocks. Each block began with a 2-second fixation period, followed by a control trial. Control trials included a 30 second video clip of fish swimming in a tank. Fish videos were shown as an affectively neutral, non-social, visuo-motor cognitive control, which we chose

to establish an affective baseline. The use of fish stimuli ensured the asociality of this condition for robust differentiation against the inductive conditions. We extracted 24 fish video clips from a continuous 8-hour video of fish movement in a tank, available online at [\[https://www.youtube.com/watch?v=6YmvqpQsBUQ\]](https://www.youtube.com/watch?v=6YmvqpQsBUQ). This video was chosen since the background, angle, and ambient sounds are consistent throughout its entirety. Control trials always preceded induction trials to minimize emotional carryover effects from the induction trials, promote a return to baseline arousal, and curb rumination. While the order of the condition presentation was consistent (i.e., fish $\rightarrow$ induction), the 24 fish and 24 induction videos were shown in a fully randomized order.

During the fish trials, the videos were presented against a grey screen with a slider bar below. The slider bar started at “0” and went up to “100” with labels on each end of the bar. Participants were instructed to select a “2” on the keyboard to move the slider to the left every time they counted a fish. If a fish left the screen and came back, they were instructed to count it again, and if they made a mistake, they were instructed to select “1” on the keyboard to slide the bar back to the right and subtract 1 from their total tally. The slider bar had a number beneath it reflecting the current entry on the continuous rating slider. When the subject made an entry on the keyboard the number would change reflecting the new tally. A step size of 1 was used for the continuous rating slider, and data was logged continuously at a rate of 24Hz. At the end of each fish video, a screen appeared with a text box asking participants to type the total number of fish they counted.

After each control trial, an induction trial was presented containing videos of either the self or other agent engaging in the Talent Show from Session 1. Each induction trial began with another 2 second fixation, followed by an audiovisual cue providing information regarding the agent of

the ensuing trial. Cues consisted of 2 second videos of either an orange circle reading “Other” paired with a 700Hz tone, or a blue circle reading “Self” paired with a 500Hz tone. Immediately after the cue, participants watched the induction videos.

Similarly to the fish trials, participants manipulated a continuous rating slider beneath the induction videos and were asked to rate the performance in a moment-to-moment fashion based on whether the video was better or worse than expected on a 7-point scale. The rating slider started in a neutral position in the middle of the slider with a rating of 4. Participants were told that lower scores indicate the video was falling short of their expectations, a score of 4 indicates the video was meeting their expectations, and a score above 4 indicates that the video was exceeding their expectations. Participants were also told not to settle on a single score, but to update their rating based on how the video struck them at a given moment. The “1” and “2” keys moved the slider to the left and right, respectively.

After each induction trial, participants were immediately asked a series of questions on a 7-point Likert scale. After every video, regardless of the agent therein, participants were asked overall if the video was better or worse than they expected and to rate the video overall. After the self-trials only, participants were asked to rate how much they would expect someone else to like their performance, and for other trials only, participants were asked to rate how much they liked the performance. Questions were always presented in the same order. Note that a condensed practice version of the task was administered before beginning the real SSOO task to ensure they understood the procedures before starting (see Appendix S1 for practice task details in Online Resource 1).

Recording Procedures

Participants were recorded continuously for the entirety of the Talent Show and were instructed to stay in the pink performance square for both performance and preparation periods. To ensure subject compliance, they were told that they would be monitored from another room. The recording camera was setup catty-corner from the performance square and was kept at a consistent zoom and angle to ensure uniformity of recording characteristics across participants. All video recordings were captured using a Tenveo PTZ Conference Camera (version v1.1U, 720p resolution, 60 fps, 114° wide field of view) and were saved locally to a designated recording laptop. Audio was recorded using a RØDE Wireless GO II Single Set wireless microphone system. The transmitter, featuring a built-in omnidirectional condenser microphone, was clipped to the participant's clothing near the chest to ensure clear vocal capture. The gain was set to 0 dB, meaning no amplification was applied to the signal. The receiver was connected to the recording laptop via a standard 3.5 mm TRS input. Audio and video were simultaneously recorded using Open Broadcaster Software (OBS Studio) throughout all experimental sessions (see Appendix S2 for details in Online Resource 1).

Video Editing Procedures

Video editing was conducted using Riverside.fm, a browser-based platform allows for collaborative post-production processing. Thirty-second clips were extracted from each session, beginning at the moment the participant initiated their performance and continuing for 30 seconds. All video material was recorded and exported at a resolution of 720 pixels vertically and 1280 pixels horizontally.

Study 1 Results

Statistical Analyses

Ratings were nested within participants, and participants were nested within dyads. We therefore used multilevel models (MLMs), conducted in RStudio version 2024.04.2 using the lme4 package (Bates et al., 2015). MLMs allowed us to account for the nonindependence of repeated trial-level ratings within participants and, where applicable, within dyads, while preserving trial-level variability rather than aggregating responses across trials. Each was addressed by a set of questions, with a separate MLM fit for each question.

To test H1 (Social PE “Liking Gap”), we fit two MLMs examining whether participants underestimated how positively another person would evaluate them. These models corresponded to two questions depicted in **Table 1** and compared actual partner ratings with participants predicted partner ratings. For Q1, participants’ Session 1 predictions, made immediately after performing, were compared with the rating another participant actually gave the performance in response to the item “How would you rate that performance overall?” The predicted-rating item was “How would you expect someone else to rate your performance?” For Q2, the same actual partner rating was compared with participants’ predictions on the same expected-rating item, but predictions were made in Session 2 after participants viewed their own performance. In both models, the contrast was interpreted as actual rating minus predicted rating; therefore, a positive effect of rating type would indicate that actual partner ratings were higher than participants predicted, supporting H1.

To test H2 (Egocentric PE “Doing vs. Viewing”), we fit two MLMs examining whether participants’ ratings shifted from immediately after performing in Session 1 to after viewing their performance in Session 2. These models correspond to two additional questions depicted in **Table 1**. Q3 compared participants’ responses to the item “How would you rate that overall?” across the two time points: after viewing in Session 2 versus immediately after performing in Session 1. Q4

used the same session-based comparison but for the item “How would you expect someone else to rate that?” In both models, the contrast was interpreted as Session 2 minus Session 1; therefore, a negative effect of session would indicate that ratings were lower after viewing than after performing, supporting H2.

For each question, we fit two versions of the model: a base model and an exploratory model. The base model included the comparison of interest, either rating type for H1 or session for H2, as well as trial order, indexed by the sequential position of each trial within the experiment. Trial order served as a proxy for elapsed time and allowed us to test whether responses changed across repeated exposure to the experiment. Because participants completed multiple trials in succession, we anticipated habituation effects, such that discrepancies would be largest early in the task and diminish over time. Models also included the interaction between the comparison of interest and trial order; significant interactions were followed up with simple slopes analyses. DF was also included as a covariate. Random intercepts were included for trial and for participant or dyad, depending on the question. The exploratory model additionally included performance modality — story, song, or dance — as an additional independent variable. This allowed us to test whether the focal effects persisted after accounting for variability in ratings across performance types, and to examine how the task design may have contributed to those effects. For each question, the exploratory model was fit three times using sequentially relevelled reference categories — story, song, and dance — to obtain all pairwise modality comparisons: song versus story, dance versus story, and dance versus song.

PE Descriptives

We computed participant-level PE scores for each question using the same directional contrasts as the MLMs. First, trial-level difference scores were calculated for each participant

and question. For Q1 and Q2, PE was defined as the actual partner rating minus the participant's predicted partner rating; positive values therefore indicate that participants underestimated how positively another person evaluated them. For Q3 and Q4, PE was defined as the Session 2 rating after viewing minus the Session 1 rating after performing; negative values therefore indicate lower ratings after viewing than after performing. Trial-level PE scores were then averaged within participant, yielding one average PE score per participant for each question (see Table S2 and **Figure S3** in Online Resource 1 for details)

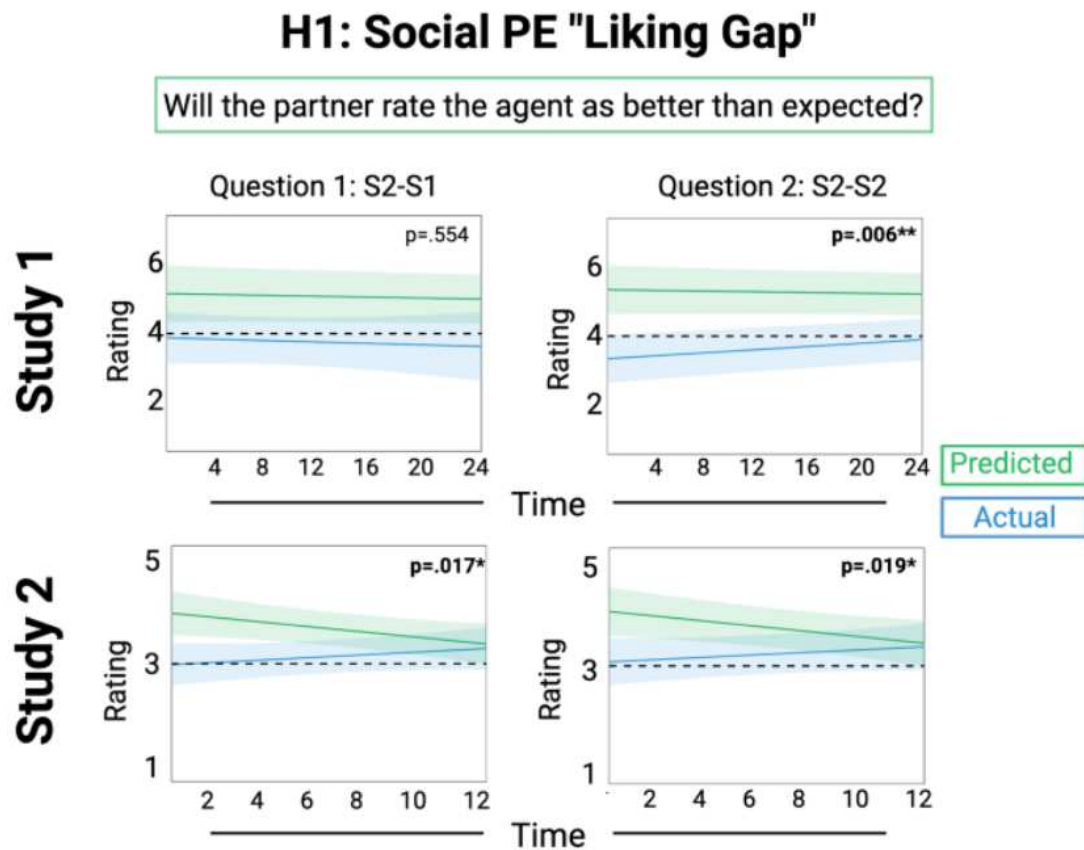


Fig. 2 Multilevel model estimates from the base models of Hypothesis 1 (H1; Social Prediction Error "Liking Gap") across Studies 1 and 2. Predicted partner ratings (green) and actual partner ratings (blue) are plotted across trial order for Question 1 (predictions made after performing in Session 1) and Question 2 (predictions made after viewing in Session 2). Shaded regions represent 95% confidence intervals. The p -values displayed in each panel correspond to the condition $\times$ trial order interaction. * $p < .05$, ** $p < .01$, *** $p < .001$.

Social Prediction Error

Participants underestimated how positively their partners would evaluate their performances (H1: Social PE, Q1). When predictions were made immediately after performing in Session 1, 75% of participants underestimated how positively their partners would rate them. In the base Q1 model, actual partner ratings were significantly higher than participants' predicted ratings ($M = 4.86$ vs. 3.66 , $p = .002$; **Figure 2**; see **Table 2** for full statistics). Ratings did not change significantly over the course of the experiment, and the gap between actual and predicted ratings remained stable over time (all p 's $> .20$). Creative DF was not significantly associated with ratings ($p = .744$). In the exploratory Q1 model accounting for performance modality, actual partner ratings remained significantly higher than predicted ratings ($p = .003$). None of the pairwise modality comparisons were significant, indicating that overall ratings did not differ across performance modalities (all $ps > .20$).

Table 2. Multilevel model results for Study 1. Each row corresponds to a predictor within a given question (Q). Base Model includes Condition, Time, Condition $\times$ Time, and Degree of Freedom as predictors. Exploratory Model additionally includes performance modality, with all pairwise comparisons reported (Dance vs. Story and Song vs. Story from the story-reference model; Dance vs. Song from the song-reference model). β = unstandardized coefficient. Bold indicates significance at $\alpha = .05$. * $p < .05$, ** $p < .01$, *** $p < .001$.

Question	Predictor	Base Model			Exploratory Model		
		β (SE)	p-value	95% CI	β (SE)	p-value	95% CI
Q1	Condition	1.12 (.35)	.002**	[.43, 1.81]	1.23 (.41)	.003**	[.43, 2.03]
	Time	-.04 (.03)	.298	[-.10, .03]	-.02 (.05)	.635	[-.11, .07]
	Condition $\times$ Time	.02 (.04)	.554	[-.05, .10]	.01 (.05)	.874	[-.09, .11]
	DF	.05 (.17)	.744	[-.27, .38]	.05 (.17)	.774	[-.28, .38]
	Modality (Dance vs. Story)	—	—	—	-.12 (.27)	.655	[-.66, .42]
	Modality (Song vs. Story)	—	—	—	-.26 (.22)	.243	[-.70, .18]
	Modality (Dance vs. Song)	—	—	—	.14 (.22)	.535	[-.30, .58]
Q2	Condition	1.99 (.31)	<.001***	[1.38, 2.60]	1.98 (.31)	<.001***	[1.38, 2.58]
	Time	.05 (.02)	<.001***	[.02, .08]	.05 (.01)	.001**	[.02, .08]
	Condition $\times$ Time	-.06 (.02)	.006**	[-.10, -.02]	-.06 (.02)	.006**	[-.10, -.02]
	DF	-.01 (.16)	.963	[-.37, .36]	-.01 (.15)	.963	[-.30, .28]
	Modality (Dance vs. Story)	—	—	—	-.45 (.18)	.014*	[-.80, -.09]

Question	Predictor	Base Model			Exploratory Model		
		β (SE)	p-value	95% CI	β (SE)	p-value	95% CI
	Modality (Song vs. Story)	—	—	—	-.48 (.18)	.008**	[-.84, -.12]
	Modality (Dance vs. Song)	—	—	—	.03 (.18)	.865	[-.32, .39]
Q3	Condition	-.87 (.38)	.023*	[-1.62, -.12]	-.62 (.40)	.122	[-1.40, .17]
	Time	.02 (.04)	.683	[-.07, .10]	.05 (.05)	.264	[-.04, .14]
	<i>Condition × Time</i>	.04 (.05)	.389	[-.05, .13]	.00 (.05)	.953	[-.09, .10]
	DF	-.07 (.31)	.833	[-.78, .64]	-.08 (.16)	.606	[-.41, .24]
	Modality (Dance vs. Story)	—	—	—	-.88 (.27)	.001**	[-1.41, -.36]
	Modality (Song vs. Story)	—	—	—	-1.09 (.22)	<.001***	[-1.52, -.66]
	Modality (Dance vs. Song)	—	—	—	.21 (.22)	.342	[-.22, .64]
Q4	Condition	-.85 (.32)	.009**	[-1.49, -.21]	-.57 (.41)	.163	[-1.38, .23]
	Time	-.01 (.03)	.860	[-.08, .06]	.02 (.05)	.605	[-.07, .12]
	<i>Condition × Time</i>	.07 (.04)	.080	[-.01, .14]	.03 (.05)	.541	[-.07, .13]
	DF	-.04 (.21)	.868	[-.53, .46]	-.05 (.17)	.788	[-.38, .28]
	Modality (Dance vs. Story)	—	—	—	-.52 (.27)	.058	[-1.06, .02]
	Modality (Song vs. Story)	—	—	—	-.74 (.22)	<.001***	[-1.18, -.30]
	Modality (Dance vs. Song)	—	—	—	.22 (.22)	.322	[-.22, .66]

332

333

334

335

336

337

338

339

340

341

342

343

The same liking gap emerged when participants made predictions after viewing their own performances in Session 2 (H1: Social PE, Q2), with 75% of participants underestimating how positively their partners would rate them. In the base Q2 model, actual partner ratings were significantly higher than participants' predictions ($M = 4.86$ vs. 3.61 , $p < .001$). Ratings increased overall across the experiment ($p < .001$), but the difference between actual and predicted partner ratings narrowed over time ($p = .006$). Simple-slopes analyses showed that participants became increasingly optimistic about how their partners would rate them ($b = 0.05$, $p < .001$), whereas actual partner ratings remained stable ($b = -0.01$, $p = .542$). Creative DF was not significantly associated with ratings ($p = .963$). In the exploratory Q2 model accounting for performance modality, the difference between actual and predicted ratings, the overall effect of time, and the narrowing of the liking gap remained significant (all $ps < .006$). Unlike in Q1,

performance modality influenced overall ratings: dance and song received lower ratings than story (both p s < .05), whereas dance and song did not differ from each other (p = .865).

Egocentric Prediction Error

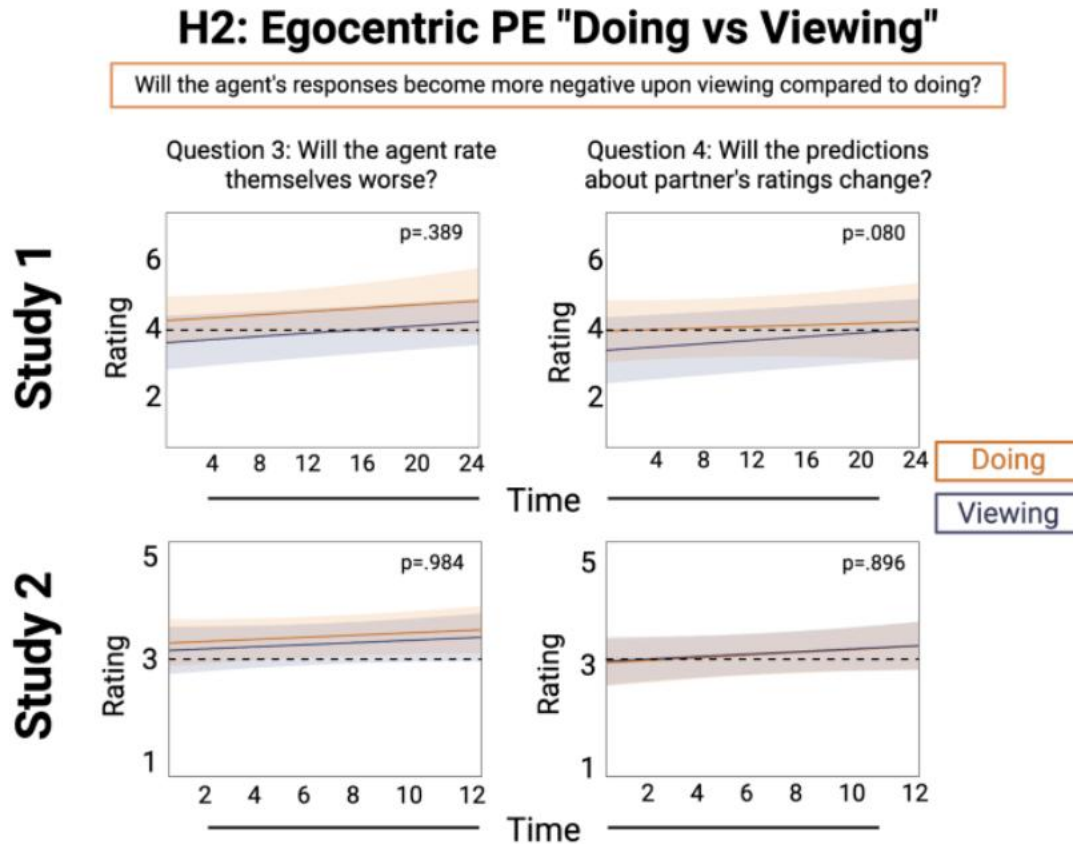


Fig. 3 Multilevel model estimates from the base models of Hypothesis 2 (H2; Egocentric Prediction Error "Doing vs. Viewing") across Studies 1 and 2. Ratings made immediately after performing (Doing; orange) and after viewing the recorded performance (Viewing; purple) are plotted across trial order for Question 3 (self-ratings) and Question 4 (predicted partner ratings). Shaded regions represent 95% confidence intervals. The p -values displayed in each panel correspond to the condition $\times$ trial order interaction. * p < .05, ** p < .01, *** p < .001.

Participants rated their performances less positively after viewing than immediately after doing them (H2: Egocentric PE, Q3). When self-ratings made immediately after performing in Session 1 ("doing") were compared with ratings made after viewing the recordings in Session 2

(“viewing”), 58.3% of participants rated themselves more negatively in the viewing condition. In the base Q3 model, self-ratings were significantly higher during doing than viewing ($M = 3.84$ vs. 3.56 , $p = .023$; **Figure 3**; see **Table 2** for full statistics). Self-ratings did not change significantly over the course of the experiment, and the difference between doing and viewing remained stable over time (all $ps > .30$). Creative DF was not significantly associated with self-ratings ($p = .833$). However, the doing-versus-viewing difference was no longer significant in the exploratory Q3 model accounting for performance modality ($p = .122$). Performance modality significantly influenced self-ratings: dance and song received lower self-ratings than story (both $ps < .001$), whereas dance and song did not differ from each other ($p = .342$).

The doing-versus-viewing analysis was also applied to participants’ predictions of how their partners would rate them (H2: Egocentric PE, Q4). When predictions made immediately after doing the performances in Session 1 were compared with predictions made after viewing them in Session 2, 41.7% of participants expected lower partner ratings in the viewing condition. In the base Q4 model, predicted partner ratings were significantly higher during doing than viewing ($M = 3.66$ vs. 3.61 , $p = .009$), suggesting that viewing their performances led participants to anticipate slightly less favorable evaluations from their partners. Predicted partner ratings did not change significantly over the course of the experiment, and the difference between doing and viewing remained stable over time (all $ps > .08$). Creative DF was not significantly associated with ratings ($p = .868$). However, the doing-versus-viewing difference was no longer significant in the exploratory Q4 model accounting for performance modality ($p = .163$). Performance modality significantly influenced predicted partner ratings: song received lower ratings than story ($p < .001$), whereas dance did not differ significantly from either story or song (both $ps > .05$).

Study 1 Discussion

Study 1 provided proof-of-concept support for the utility of PGT as a paradigm for eliciting socio-affective PEs (PEs). Consistent with H1 (Social PE "Liking Gap"), participants systematically underestimated how positively their partner would evaluate their performance, replicating the liking gap effect previously observed in naturalistic social interactions (Boothby et al., 2018). Importantly, this effect emerged both immediately after performing and after participants viewed recordings of their own performances, suggesting that the tendency to underestimate others' ratings is robust across multiple stages of the task. This suggests that the task successfully created a socially evaluative environment capable of inducing social PEs. Additionally, there is evidence that participants habituated during the SSOO. The liking gap measured in Session 2 during the SSOO closed over time. This was driven by increasingly positive predictions about how partners would evaluate participants, while actual partner ratings remained stable. This pattern is consistent with internal model updating, whereby repeated exposure to their performances may have gradually reduced PEs until expectations were no longer violated. However, predictions made immediately after performing in Session 1 remained stable over time, perhaps because there are differences in the resources available for self-referential processing when performing a task versus passively viewing a task.

Consistent with H2 (Egocentric PE "Doing vs. Viewing"), the base models suggested that participants rated both themselves and their expected partner ratings more positively immediately after performing than after viewing their recordings. This indicates that our embarrassment induction successfully created an egocentric PE, such that participants expected their performances to be better than they actually judged them to be after viewing. However, this effect did not emerge in the exploratory model, suggesting it may be less robust than the liking

gap and more sensitive to differences across performance modalities. This finding suggests that simply viewing one's own performance may not be sufficient to produce robust egocentric PEs.

Contrary to our subsidiary hypothesis, the degrees of freedom manipulation did not meaningfully influence PE magnitude. Performance modality, however, accounted for significant variance across several outcomes. Story performances were generally rated more positively than songs and dances, suggesting that different performance types may vary in their social consequences.

Study 2

The goal of Study 2 was to conceptually replicate and extend the findings of Study 1 in a larger sample while correcting some methodological limitations. First, we added a tightly matched control condition, allowing us to isolate the social norm violation from other aspects of the task, such as the effect of self- versus other viewing. Second, the DF manipulation was eliminated given the null results in Study 1. Third, we eliminated the seated observer next to the participant (included to provide a salient social context), because this limits the utility of our task in, for instance, fMRI studies. Rather, we asked whether simply knowing that an observer was present in an adjacent room would be sufficient to elicit the effect. Fourth, we randomized trials in Study 2 so that we could disentangle habituation from order-specific effects. Last, in Study 2 we explicitly measured embarrassment to gain insight into what participants experienced emotionally. In other regards, Study 2 closely matched Study 1. The methods below focus only on the procedural and task changes introduced in Study 2.

Accordingly, Study 2 tested the same core hypotheses as Study 1, examining whether the H1 (Social PE "Liking Gap") and H2 (Egocentric PE "Doing vs. Viewing") would replicate in a larger sample using the modified paradigm. In addition, a third hypothesis, H3 (Affective PE "Self

vs Other”) was introduced to test whether participants would report more negative emotional valence when viewing their own performance than when viewing another person's performance (Table 1). Passive control conditions were included under H1 and H3 to determine whether these effects were specific to embarrassment-induction trials. These conditions were not hypothesized to elicit PEs. Finally, exploratory analyses examined the affective correlates of PEs, including whether PE magnitude predicted emotion endorsement and whether emotion endorsements differed between self and other induction trials.

Study 2 Methods

Participants

We recruited 28 participants. Five were lost due to recording or other technical issues, and two were lost to attrition, resulting in a final sample of 20 participants. Participants were recruited and compensated in the same manner described in Study 1 (See Table S1 for demographics in Online Resource 1).

Session 1: Changes to the Talent Show

The overall structure of the Talent Show remained the same, with several methodological improvements. First, participants selected at least four out of eight possible tasks based on prior familiarity, allowing us to assign dyads identical task sets (see Figure S1 in Online Resource 1). Second, prior to the Talent Show, we recorded in a three-quarter view while participants completed the self-report questionnaires. This addition aimed to generate *passive trials* as a stimulus that could later be used as a tightly matched control trial to isolate emotion induction from neutral self or other processing. Third, we removed the manipulation of creative freedom (i.e., high vs. low DF) from the Talent Show after finding no significant behavioral differences in Study 1. To maintain the creativity study deception, participants were instead instructed to put

their own creative flair into each performance. Fourth, we introduced a new creative modality in which participants were asked to rap children’s nursery rhymes. In sum, participants completed three trials in each of four modalities—storytelling, singing, rapping a nursery rhyme, and dancing—for a total of twelve performance trials. Fifth, in this iteration participants were given two minutes to prepare for every performance, during which they received a visual cue corresponding to their assigned prompt—such as song lyrics, a looping dance clip, a storyboard, or the text of a nursery rhyme. Unlike Study 1, in which prompts were explained verbally by the experimenter, participants in Study 2 received a persistent visual cue throughout the preparation period to reduce memory demands and promote consistency in performance content across participants. After each preparatory period, subjects completed their one-minute performance. Sixth, the scale used to rate their own performance and estimate partner ratings was reduced from 7 points to 5 points, as responses in Study 1 clustered around the midpoint with extreme values rarely used (**Figure 5**, Panel A; see Appendix S3 in Online Resource 1). Seventh, trial order was randomized to prevent order and modality from confounding the results.

Finally, upon completing the Talent Show, participants filled out a post-task emotion survey prompting them to reflect on their affective experiences during each performance modality (e.g., storytelling, singing, etc.; see **Figure S2** for a task schematic in Online Resource 1). For each modality, participants were asked to select a single emotion that best described how they felt while performing, choosing from the following options: amused, angry, bored, embarrassed, or proud. Once an emotion was selected, they rated the intensity of that emotion on a 5-point scale.

Session 2: Changes to the SSOO

The overall structure of the SSOO remained the same, with several methodological changes. First, instead of an in-person observer, participants were informed that their screen would

be monitored in real time by an unfamiliar experimenter to increase perceived social discomfort. Second, participants viewed 30-second video clips of their own recorded performances (12 self-induction trials), their partner's performances (12 other-induction trials), passive recordings of themselves and their partner completing surveys (6 self-passive and 6 other-passive trials), and 12 neutral fish videos. While viewing the videos, participants rated the stimuli in real time, as they had in Study 1. For both passive and induction clips, participants used a continuous scale ranging from 1–5 to track and update their emotional experience. All trials were presented in a pseudo-randomized order, with no more than two consecutive trials of the same type (fish, self, or other). Third, each video was preceded by an orange circular cue labeled “fish,” “self,” or “other” and paired with a 500 Hz tone to reduce surprise and help participants anticipate the upcoming content. To further support identification, an icon appeared in the upper-right corner of each cue: a fish for fish trials, a notebook for passive trials, and a microphone for induction trials. Trials were separated by a jittered inter-trial interval (ITI) of 2, 4, or 6 seconds.

Fourth, after each video, the same three questions were presented after every trial type—including fish trials. Using the same 5-point Likert scale from Session 1, participants rated (1) the performance itself, (2) how they expected someone else would rate it, and (3) how positively or negatively they felt while watching the performance (see Appendix S3 in Online Resource 1). A 4-second response window was implemented for each question, during which participants could revise their entry, with only the final selection recorded.

Fifth, participants completed a post-task emotion survey designed to assess their emotional experience during the session. For each trial type, participants were shown a representative screenshot (one image per creative modality per agent, one passive per agent, and one fish), along with text indicating whether the trial featured 'self' or 'other' and the corresponding modality. As

in Session 1, participants selected the emotion that best described their experience and rated its intensity.

Study 2 Results

Statistical Analyses

The same MLM approach described in Study 1 was used throughout, with a separate model fit for each question. As in Study 1, Q1 and Q2 tested H1 (Social PE "Liking Gap") by comparing actual partner ratings with participants' predicted partner ratings. Q3 and Q4 tested H2 (Egocentric PE "Doing vs. Viewing") by comparing ratings made immediately after performing in Session 1 with ratings made after viewing one's own performance in Session 2. These models used the same contrast directions and model structures described in Study 1.

Study 2 additionally examined emotional valence differences. To test H3 (Affective PE "Self vs Other"), Q5 examined participants' responses to the item, "How positive or negative did you feel about that performance?", for self versus other performances. The model included condition (self vs. other) and trial order as fixed effects, with random intercepts for participant and trial index. The condition contrast was interpreted as other minus self; therefore, a significant positive effect of condition would indicate that participants evaluated their partner's performance more positively than their own, supporting H3.

Q6 and Q7 served as passive-control analogues of the primary PE questions. Q6 mirrored the H1 (Social PE "Liking Gap") comparison by testing whether actual and predicted ratings differed during passive viewing trials. Q7 mirrored the H3 (Affective PE "Self vs Other") comparison by testing whether ratings differed between self and other passive-viewing trials. Because passive trials did not involve the induction context or norm-violation component present in the faux pas condition, we expected little to no discrepancy in either model. Thus, effects near

zero in Q6 and Q7 would be consistent with the prediction that PE-like discrepancies are specific to induction trials.

For Study 2, model structure, random effects, and modality-adjustment procedures were otherwise identical to those described in Study 1. Base and exploratory models were fit for each question, with exploratory models relevelled to obtain all pairwise comparisons among story, song, rhyme, and dance performances.

PE Descriptives

Participant-level PE scores were computed using the same approach described in Study 1. First, trial-level difference scores were calculated for each participant and question and then averaged within participant, yielding one average PE score per participant for each question. For Q1–Q4, PE scores were computed using the same directional contrasts described in Study 1. For Q5, PE was defined as other rating minus self-rating, such that positive values indicate more favorable ratings of the partner’s performance relative to one’s own. For the passive-control questions, Q6 and Q7, values close to zero were expected, reflecting little to no discrepancy in the absence of the induction context (see **Table S2** and **Figure S3** in Online Resource 1).

Emotion Endorsement Analysis

To compare the proportions of amused and embarrassed endorsements between the self and other conditions, we conducted an exploratory Fisher’s exact test on a 2×2 contingency table. Analyses were restricted to these two emotions because they were the most frequently endorsed across participants. Fisher’s exact test was used instead of a chi-square test because several cells contained small, expected frequencies, violating chi-squared assumptions. To further examine whether PEs were associated with emotional responses to the induction, we fit a series of logistic mixed-effects models predicting emotion endorsement from PE difference

scores. For each question, a separate model was fit with PE difference score as the predictor and embarrassment endorsement (embarrassed = 1, amused = 0) as the outcome, with random intercepts for participants.

Social Prediction Error

Participants underestimated how positively their partners would evaluate their performances (H1: Social PE, Q1). When predictions were made immediately after performing in Session 1, 75% of participants underestimated how positively their partners would rate them. In the base Q1 model, actual partner ratings were significantly higher than participants' predicted ratings ($M = 3.63$ vs. 3.11 , $p < .001$; **Figure 2**; see **Table 3** for full statistics). The difference between actual and predicted partner ratings narrowed over time ($p = .017$). Simple-slopes analyses showed that actual partner ratings declined across the experiment ($b = -0.05$, $p = .028$), whereas participants' predicted ratings remained stable ($b = 0.03$, $p = .241$). Thus, the liking gap narrowed as actual partner ratings approached participants' predictions. In the exploratory Q1 model accounting for performance modality, both the difference between actual and predicted ratings and the narrowing of the liking gap over time remained significant (both $ps < .012$). Performance modality also influenced overall ratings: dance received higher ratings than story, song, and rhyme, whereas rhyme received lower ratings than story and song (all relevant $ps \leq .033$).

Table 3. Multilevel model results for Study 2. Each row corresponds to a predictor within a given question (Q). Base Model includes Condition, Time, and Condition $\times$ Time as predictors. Exploratory Model additionally includes performance modality, with all pairwise comparisons reported (Song vs. Story, Dance vs. Story, and Rhyme vs. Story from the story-reference model; Dance vs. Song and Rhyme vs. Song from the song-reference model; Rhyme vs. Dance from the dance-reference model). Questions 6 and 7 are passive control questions with no exploratory model. β = unstandardized coefficient. Bold indicates significance at $\alpha = .05$. * $p < .05$, ** $p < .01$, *** $p < .001$.

Question	Predictor	Base Model			Exploratory Model		
		β (SE)	p-value	95% CI	β (SE)	p-value	95% CI
Q1	Condition	1.02 (.23)	<.001***	[.56, 1.47]	1.04 (.23)	<.001***	[.58, 1.49]
	Time	.03 (.02)	.240	[-.02, .07]	.03 (.02)	.219	[-.02, .07]
	<i>Condition × Time</i>	-.08 (.03)	.017*	[-.14, -.01]	-.08 (.03)	.012*	[-.14, -.02]
	Modality (Song vs. Story)	—	—	—	-.06 (.15)	.701	[-.36, .24]
	Modality (Dance vs. Story)	—	—	—	.36 (.15)	.019*	[.06, .67]
	Modality (Rhyme vs. Story)	—	—	—	-.39 (.16)	.013*	[-.69, -.08]
	Modality (Dance vs. Song)	—	—	—	.42 (.15)	.006**	[.12, .72]
	Modality (Rhyme vs. Song)	—	—	—	-.33 (.15)	.033*	[-.63, -.03]
	Modality (Rhyme vs. Dance)	—	—	—	-.75 (.15)	<.001***	[-1.06, -.45]
Q2	Condition	.97 (.23)	<.001***	[.53, 1.41]	1.02 (.23)	<.001***	[.57, 1.46]
	Time	.02 (.02)	.250	[-.02, .07]	.03 (.02)	.240	[-.02, .07]
	<i>Condition × Time</i>	-.07 (.03)	.019*	[-.13, -.01]	-.08 (.03)	.010*	[-.14, -.02]
	Modality (Song vs. Story)	—	—	—	-.12 (.15)	.438	[-.41, .18]
	Modality (Dance vs. Story)	—	—	—	.22 (.15)	.146	[-.08, .51]
	Modality (Rhyme vs. Story)	—	—	—	-.44 (.15)	.004**	[-.74, -.15]
	Modality (Dance vs. Song)	—	—	—	.33 (.15)	.025*	[.04, .62]
	Modality (Rhyme vs. Song)	—	—	—	-.33 (.15)	.029*	[-.62, -.03]
	Modality (Rhyme vs. Dance)	—	—	—	-.66 (.15)	<.001***	[-.96, -.37]
Q3	Condition	-.14 (.22)	.509	[-.57, .28]	-.14 (.22)	.525	[-.57, .29]
	Time	.02 (.02)	.309	[-.02, .06]	.02 (.02)	.282	[-.02, .06]
	<i>Condition × Time</i>	.00 (.03)	.984	[-.06, .06]	-.00 (.03)	.997	[-.06, .06]
	Modality (Song vs. Story)	—	—	—	-.12 (.14)	.405	[-.40, .16]
	Modality (Dance vs. Story)	—	—	—	.63 (.14)	<.001***	[.34, .91]
	Modality (Rhyme vs. Story)	—	—	—	-.16 (.14)	.272	[-.44, .13]
	Modality (Dance vs. Song)	—	—	—	.75 (.14)	<.001***	[.47, 1.03]
	Modality (Rhyme vs. Song)	—	—	—	-.04 (.14)	.786	[-.32, .24]
	Modality (Rhyme vs. Dance)	—	—	—	-.79 (.14)	<.001***	[-1.07, -.50]
Q4	Condition	.05 (.22)	.840	[-.39, .48]	.03 (.22)	.877	[-.41, .48]
	Time	.03 (.02)	.191	[-.01, .07]	.03 (.02)	.186	[-.01, .07]
	<i>Condition × Time</i>	-.00 (.03)	.896	[-.06, .06]	-.00 (.03)	.936	[-.06, .06]
	Modality (Song vs. Story)	—	—	—	-.10 (.17)	.572	[-.48, .28]
	Modality (Dance vs. Story)	—	—	—	.50 (.17)	.015*	[.12, .88]
	Modality (Rhyme vs. Story)	—	—	—	-.32 (.17)	.090	[-.70, .06]
	Modality (Dance vs. Song)	—	—	—	.60 (.17)	.006**	[.22, .98]
	Modality (Rhyme vs. Song)	—	—	—	-.22 (.17)	.226	[-.61, .16]
	Modality (Rhyme vs. Dance)	—	—	—	-.82 (.17)	<.001***	[-1.20, -.44]
Q5	Condition	-.77 (.20)	<.001***	[-1.16, -.39]	-.79 (.20)	<.001***	[-1.17, -.40]
	Time	-.06 (.02)	.002**	[-.10, -.02]	-.06 (.02)	.001**	[-.10, -.02]
	<i>Condition × Time</i>	.07 (.03)	.013*	[.01, .12]	.07 (.03)	.011*	[.02, .12]
	Modality (Song vs. Story)	—	—	—	.07 (.13)	.633	[-.23, .36]
	Modality (Dance vs. Story)	—	—	—	.30 (.13)	.045*	[.01, .59]
	Modality (Rhyme vs. Story)	—	—	—	-.03 (.13)	.812	[-.33, .26]
	Modality (Dance vs. Song)	—	—	—	.23 (.13)	.105	[-.06, .53]
	Modality (Rhyme vs. Song)	—	—	—	-.10 (.13)	.478	[-.39, .20]
	Modality (Rhyme vs. Dance)	—	—	—	-.33 (.13)	.030*	[-.62, -.04]
Q6	Condition	.13 (.18)	.462	[-.22, .47]	—	—	—
	Time	-.00 (.03)	.883	[-.07, .06]	—	—	—
	<i>Condition × Time</i>	-.02 (.04)	.628	[-.11, .07]	—	—	—
Q7	Condition	-.48 (.19)	.011*	[-.85, -.11]	—	—	—
	Time	-.06 (.03)	.057	[-.13, .00]	—	—	—
	<i>Condition × Time</i>	.11 (.05)	.022*	[.02, .21]	—	—	—

The same liking gap emerged when participants made predictions after viewing their own performances in Session 2 (H1: Social PE, Q2), with 65% of participants underestimating how positively their partners would rate them. In the base Q2 model, actual partner ratings were significantly higher than participants' predicted ratings ($M = 3.63$ vs. 3.14 , $p < .001$). The difference between actual and predicted ratings narrowed over time ($p = .019$). Simple-slopes analyses showed that actual partner ratings declined across the experiment ($b = -0.05$, $p = .031$), whereas participants' predicted ratings remained stable ($b = 0.02$, $p = .251$). In the exploratory Q2 model accounting for performance modality, both the difference between actual and predicted ratings and the narrowing of the liking gap remained significant (both $ps < .010$). Performance modality also influenced overall ratings: dance received higher ratings than song and rhyme (both $ps < .025$), whereas rhyme received lower ratings than story and song (both $ps < .029$).

Egocentric Prediction Error

Participants did not reliably rate their performances more negatively or predict that their partners would rate them more negatively after viewing than immediately after performing (H2: Egocentric PE, Q3–Q4). When ratings made immediately after performing in Session 1 (“doing”) were compared with ratings made after viewing the recordings in Session 2 (“viewing”), 60% of participants rated themselves more negatively after viewing, whereas 45% predicted that their partners would rate them more negatively. However, in the base models, neither self-ratings nor predicted partner ratings differed significantly between doing and viewing. Neither outcome changed significantly over time, and the doing-versus-viewing differences remained stable across the experiment (all $ps > .19$; **Figure 3**; see **Table 3** for full statistics). These differences also remained nonsignificant in the exploratory models accounting

for performance modality (all $ps > .52$). Performance modality nevertheless influenced both outcomes: dance received higher self-ratings than story, rhyme, and song (all $ps < .001$) and higher predicted partner ratings than story, song, and rhyme (all $ps < .015$).

Affective Prediction Error

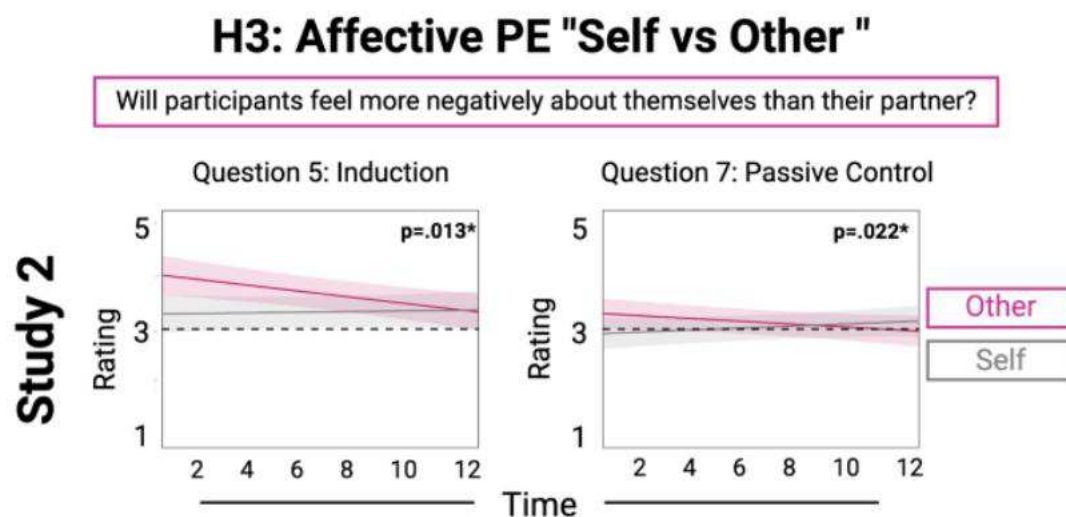


Fig. 4. Multilevel model estimates from the base models of Hypothesis 3 (H3; Affective Prediction Error "Self vs. Other") in Study 2. Ratings for self (grey) and other (pink) are plotted across trial order for Question 5 (induction trials) and Question 7 (passive control trials). Shaded regions represent 95% confidence intervals. The p -values displayed in each panel correspond to the condition $\times$ trial order interaction. * $p < .05$, ** $p < .01$, *** $p < .001$.

Participants reported feeling more negatively about their own performances than about their partners' performances (H3: Affective PE, Q5). Overall, 60% of participants reported feeling more negatively when viewing themselves than when viewing their partners. In the base Q5 model, participants reported significantly more positive feelings toward their partners' performances than toward their own ($M = 3.74$ vs. 3.24 , $p < .001$; **Figure 4**; see **Table 3** for full statistics). Reported feelings became more negative overall across the experiment, and the difference between feelings toward participants' own and their partners' performances narrowed

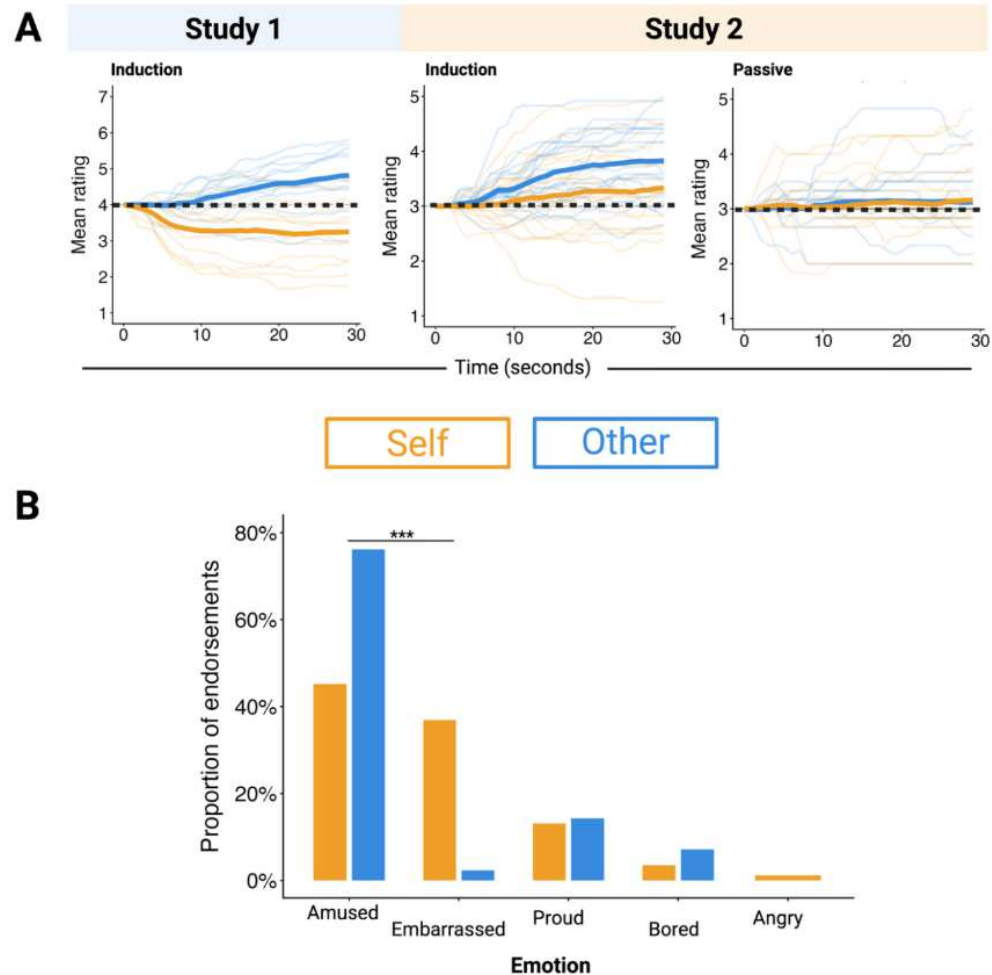
over time (both $ps < .013$). Simple-slopes analyses showed that reported feelings toward partners' performances became more negative across trials ($b = -0.06, p = .002$), whereas reported feelings toward participants' own performances remained stable ($b = 0.01, p = .737$). Thus, the self–other difference narrowed as feelings toward partners' performances became more similar to feelings toward participants' own performances. These effects remained significant in the exploratory Q5 model accounting for performance modality (all $ps < .011$). Performance modality also influenced reported feelings: dance elicited more positive feelings than story and rhyme (both $ps < .045$).

The passive control analyses examined whether the liking gap and self–other valence difference were specific to performance trials. During passive viewing, actual and predicted partner ratings did not differ significantly (H1: Social PE, Q6). Ratings did not change over time, and the difference between actual and predicted ratings remained stable across trials (all $ps > .628$). Actual and predicted ratings were also similar in magnitude ($M = 3.13$ vs. 3.08), indicating that the liking gap did not emerge during passive viewing. In contrast, participants reported different levels of valence when passively viewing themselves and their partners (H3: Affective PE, Q7). In the base Q7 model, self- and other-directed valence differed significantly, and this difference narrowed across trials (both $ps < .022$). Simple-slopes analyses showed a marginal decline in other-directed valence over time ($b = -0.06, p = .058$), whereas self-directed valence remained stable ($b = 0.05, p = .176$).

Emotion Endorsement

Participants reported feeling embarrassed much more often when viewing their own performances than when viewing their partners' performances ($n = 27$ vs. 2). In contrast, they reported feeling amused more often when viewing their partners than when viewing themselves

634 ($n = 60$ vs. 38; **Figure 5**, Panel B). Accordingly, the odds of reporting embarrassment rather than
 635 amusement were approximately 21 times higher in the self condition than in the other condition
 636 ($p < .001$).



637

638 **Fig. 5 (A)** Continuous ratings for self (orange) and other (blue) performance videos across
 639 Studies 1 and 2. The dashed horizontal line corresponds to performance meeting expectations;
 640 ratings above the line indicate the performance was better than expected, whereas ratings below
 641 indicate the performance was worse than expected. Each thin line is one participants' mean
 642 ratings, binned into 1-second intervals. Thick lines represent the group mean. **(B)** Proportion of
 643 emotion endorsements following the Surprise Self-Other Observation task (SSOO) in Study 2.
 644 Each bar represents the proportion of participants selecting the emotion that best described their
 645 experience while viewing performance videos. The asterisk denotes the result of a Fisher's exact
 646 test comparing the proportion of amused endorsements between conditions. *** $p < .001$.

We next examined whether larger PEs were associated with a greater likelihood of reporting embarrassment rather than amusement. Across the H1 and H2 questions and both passive control questions, PE magnitude was not significantly associated with reported emotion (all $ps > .165$; see **Table 4** for full statistics). The only significant association emerged for the self–other valence difference (H3: Affective PE, Q5). Participants who reported feeling more negatively about their own performances relative to their partners’ performances were more likely to report embarrassment rather than amusement ($p = .028$). This association remained significant after accounting for performance modality ($p = .026$).

Table 4. Logistic regression results predicting embarrassment endorsement (vs. amusement) from prediction error scores. Base Model includes prediction error score only with a random intercept for participant. Exploratory Model additionally includes performance modality, with all pairwise comparisons reported (Dance vs. Story, Rhyme vs. Story, and Song vs. Story from the story-reference model; Dance vs. Song and Rhyme vs. Song from the song-reference model; Rhyme vs. Dance from the dance-reference model). Q6 and Q7 are passive control questions with no exploratory model. β = unstandardized log-odds coefficient. Bold indicates significance at $\alpha = .05$. * $p < .05$, ** $p < .01$, *** $p < .001$.

Question	Predictor	Base Model			Exploratory Model		
		β (SE)	z	p	β (SE)	z	p
Q1	PE	.54 (.50)	1.08	.280	.61 (.55)	1.12	.265
	Modality (Dance vs. Story)	—	—	—	-1.60 (.97)	-1.65	.099
	Modality (Rhyme vs. Story)	—	—	—	.36 (.84)	.43	.669
	Modality (Song vs. Story)	—	—	—	-.03 (.83)	-.04	.967
	Modality (Dance vs. Song)	—	—	—	-1.57 (.93)	-1.68	.092
	Modality (Rhyme vs. Song)	—	—	—	.40 (.81)	.49	.627
	Modality (Rhyme vs. Dance)	—	—	—	1.96 (.96)	2.06	.040*
Q2	PE	.57 (.41)	1.39	.165	.60 (.45)	1.34	.181
	Modality (Dance vs. Story)	—	—	—	-1.57 (.97)	-1.61	.107
	Modality (Rhyme vs. Story)	—	—	—	.34 (.84)	.41	.683
	Modality (Song vs. Story)	—	—	—	.02 (.84)	.02	.984
	Modality (Dance vs. Song)	—	—	—	-1.59 (.94)	-1.70	.090
	Modality (Rhyme vs. Song)	—	—	—	.33 (.81)	.40	.688
	Modality (Rhyme vs. Dance)	—	—	—	1.91 (.95)	2.01	.045*
Q3	PE	-.34 (.50)	-.68	.497	-.31 (.54)	-.57	.570
	Modality (Dance vs. Story)	—	—	—	-1.60 (.98)	-1.64	.101
	Modality (Rhyme vs. Story)	—	—	—	.30 (.84)	.36	.719
	Modality (Song vs. Story)	—	—	—	-.03 (.84)	-.04	.971
	Modality (Dance vs. Song)	—	—	—	-1.57 (.93)	-1.68	.092
	Modality (Rhyme vs. Song)	—	—	—	.33 (.81)	.41	.681
	Modality (Rhyme vs. Dance)	—	—	—	1.90 (.95)	2.00	.046*
Q4	PE	-.35 (.56)	-.62	.534	-.33 (.61)	-.54	.590
	Modality (Dance vs. Story)	—	—	—	-1.62 (.98)	-1.66	.097
	Modality (Rhyme vs. Story)	—	—	—	.28 (.84)	.34	.735
	Modality (Song vs. Story)	—	—	—	-.04 (.84)	-.05	.958

	Modality (Dance vs. Song)	—	—	—	-1.58 (.94)	-1.69	.091
	Modality (Rhyme vs. Song)	—	—	—	.33 (.81)	.40	.686
	Modality (Rhyme vs. Dance)	—	—	—	1.91 (.96)	2.00	.046*
Q5	PE	.87 (.40)	2.20	.028*	.97 (.44)	2.23	.026*
	Modality (Dance vs. Story)	—	—	—	-1.66 (1.00)	-1.67	.094
	Modality (Rhyme vs. Story)	—	—	—	.34 (.83)	.41	.684
	Modality (Song vs. Story)	—	—	—	.06 (.83)	.07	.944
	Modality (Dance vs. Song)	—	—	—	-1.72 (.96)	-1.79	.073
	Modality (Rhyme vs. Song)	—	—	—	.28 (.79)	.35	.726
	Modality (Rhyme vs. Dance)	—	—	—	2.00 (.97)	2.06	.040*
Q6	PE	.84 (.69)	1.21	.225	—	—	—
Q7	PE	.44 (.80)	.55	.581	—	—	—

Study 2 Discussion

Study 2 replicated and extended the findings of Study 1. Consistent with H1 (Social PE "Liking Gap"), participants again underestimated how positively they would be evaluated by others, demonstrating a robust liking gap whether predictions were made immediately after performing or after viewing their own recordings. Importantly, this effect was absent in the passive control condition, indicating that the liking gap depends on the social-evaluative context created by the induction rather than emerging during passive viewing alone. H2 (Egocentric PE "Doing vs. Viewing") was not supported. Participants did not evaluate themselves more negatively after viewing their performances, nor did they revise their expectations of how others would evaluate them.

Study 2 additionally revealed a reliable self–other valence difference. Consistent with H3 (Affective PE "Self vs Other"), participants experienced their own performances more negatively than their partner's performances. Although a smaller self–other valence difference remained present in the passive control condition, the effect was substantially stronger during induction trials, suggesting that the norm-violating context amplified rather than created this asymmetry.

A notable secondary finding was that both the liking gap and the self–other valence difference diminished across trials. In both cases, this reduction was driven primarily by changes

in ratings of others rather than changes in self-ratings, suggesting that repeated exposure to norm-violating performances may have reduced the salience of the violations themselves. Additionally, performance modality influenced several outcomes. Unlike Study 1, dance performances were generally associated with more positive ratings and affective responses than other performance types, whereas rhyme performances often elicited less favorable ratings.

The emotion endorsement results further clarify the nature of the socio-affective processes elicited by the paradigm. Participants were substantially more likely to endorse embarrassment when evaluating their own performances and amusement when evaluating their partner's performances. Moreover, the self–other valence difference was the only PE measure associated with embarrassment endorsement. Neither the liking gap nor the doing-versus-viewing measures predicted emotional responses.

General Discussion

In the present work, we introduce and test a novel experimental paradigm called Philadelphia's Got Talent, designed to elicit socio-affective PEs. Our goal was to improve the ecological validity and temporal fidelity of prior tasks used in this realm. Across two studies, we induced socio-affective PEs by having participants view and evaluate both their own performance and the performance of a partner under surprise, norm-violating conditions. By comparing self-ratings, ratings of others, and predictions of others' ratings, the paradigm captured mismatches between expected and actual socio-affective outcomes in a naturalistic setting.

This study yielded several findings. First, participants consistently underestimated how positively another person would evaluate their performance, replicating the liking gap effect (Boothby et al., 2018). Second, participants rated their own embarrassing videos more negatively than their partner's videos on the emotional valence measure. On post-task emotion ratings

participants often endorsed “embarrassment” for themselves but not for their partner. Instead, they reported that they felt amusement when viewing other people’s performance videos. Third, there was some habituation across time, indicating that internal models were being updated.

The Liking Gap

The liking gap has primarily been studied in socially evaluative settings, such as interpersonal conversations (Boothby et al., 2018). Our findings suggest that social evaluation alone may be insufficient to produce the effect. The liking gap emerged in the induction conditions but not in the passive condition, even though both involved the possibility of social evaluation. This pattern suggests that an additional feature of the induction conditions may be necessary. One possibility is that these conditions created a self-presentational predicament—a situation in which people become concerned that their behavior may threaten the impression they want others to form of them (Leary et al., 1996). More vulnerable contexts, such as performing in front of others or generating and expressing ideas during conversation, may therefore be especially likely to produce the liking gap because they make aspects of the “self” available for others’ evaluation. This pattern also supports the validity of the passive control condition as a well-matched baseline that preserved the visual and social features of the induction while eliminating the potential faux pas.

Emotion Ratings: Flipping Self Enhancement Effects

Participants rated their own embarrassing videos more negatively than their partner’s videos on the emotional valence measure. This effect was much smaller in passive conditions. Post-task emotion ratings converged with this pattern, demonstrating that participants often endorsed “embarrassment” for themselves but not for their partner. Together, these findings support the conclusion that viewing one’s own embarrassment-induction videos elicited a negatively valenced affective experience associated with negative PEs and self-reported feelings

of embarrassment. This finding is particularly notable considering the extensive literature on self-enhancement, which refers to the tendency for individuals to maintain unrealistically positive views of themselves (Alicke & Sedikides, 2009). For example, people overestimate their own performance, predict they will behave more morally than others, and rate their performance, abilities, and traits more favorably than those of the average peer (Dunning, 2011; Epley & Dunning, 2000; Kim et al., 2010; Sedikides & Gregg, 2003; Shepperd et al., 2013; Taylor & Brown, 1988). Our findings indicate that this otherwise robust tendency can be reversed when social norms are being violated. Participants consistently evaluated themselves more negatively than their partner, a pattern that was substantially amplified during embarrassment-induction trials. Thus, although a modest self–other asymmetry was present during passive viewing, the social norm violation appears to intensify this tendency and increase the associated self-conscious affect.

Furthermore, this effect is noteworthy in light of embarrassment paradigms that attempt to induce embarrassment vicariously, such as by having participants watch embarrassing videos of others or read embarrassing vignettes. Our findings suggest that vicarious embarrassment will produce weaker or even different emotional effects as compared to firsthand embarrassment. Because firsthand, or egocentric embarrassment directs attention toward the self and how one is being perceived, it may temporarily leave fewer cognitive resources available for understanding another person’s experience (Leary et al., 1996).

Habituation occurred, indicating that internal models were updated

Across studies, evidence for habituation emerged for the liking-gap and self–other valence effects, both of which tended to diminish over the course of the task. These reductions were driven primarily by changes in evaluations of others rather than changes in self-ratings, suggesting that repeated exposure to norm-violating performances may have reduced the salience of the violations

themselves. This pattern is broadly consistent with predictive-processing accounts proposing that PEs decrease as uncertainty is reduced and internal models are updated through repeated exposure (Friston, 2010; Joiner et al., 2017). However, habituation was not observed uniformly across all measures, indicating that some socio-affective responses may be more resistant to updating than others. Nevertheless, the repeated observation of narrowing discrepancies supports our characterization of the paradigm as a relatively “top-heavy” task, in which the strongest effects tend to emerge early in the session. This temporal dynamic may itself be informative for future work. Depending on the research question, habituation could be treated either as a phenomenon of theoretical interest or as a source of variance to be modeled statistically.

Effect of Audience Presence

The results of Study 1 suggested that participants evaluated both themselves and their predicted partner evaluations more negatively immediately after viewing their own performance. However, this egocentric PE did not emerge in Study 2. One possible explanation for the discrepancy between studies is that this egocentric PE is more sensitive to characteristics of the social context. Participants in Study 1 completed the viewing phase with an experimenter physically present, whereas participants in Study 2 were alone and informed that an experimenter was observing from another room. Audience effects are known to play a central role in embarrassment (Tangney et al., 1996; Dahl et al., 2001; Crozier, 2014), and embarrassment is probably more intense when others are physically present (Riddell et al., 2025; Shearn et al., 1992; Keltner et al., 1997). It is also possible that individual differences played some role. Across analyses, only 45–60% of participants exhibited effects in the hypothesized direction, suggesting substantial heterogeneity in responses. Given our sample size, we could not delve into

these variables. Future work should aim to identify the conditions under which egocentric PEs emerge.

Prediction Errors and Self-Conscious Emotions

The present findings provide initial explicit support for embarrassment as a useful model for studying socio-affective PEs. Although embarrassment may be especially well suited for studying rapid, self-relevant socio-affective PEs, it is unlikely to be the only emotion associated with them. A substantial body of theoretical and empirical work suggests that emotional reactions are shaped, at least in part, by the extent to which outcomes deviate from expectations (Bhatia et al., 2019; Shepperd & McNulty, 2002; Villano et al., 2020). From this perspective, prediction errors are not only learning signals but also affective events, with the resulting emotion depending on how the discrepancy is appraised. Other self-conscious emotions, including shame, guilt, and pride, may therefore also arise in response to socio-affective PEs, although they differ in their underlying appraisal dimensions (Tangney & Dearing, 2002; Tracy & Robins, 2004). Guilt centers on a specific action and promotes repair (Baumeister et al., 1994), shame reflects a broader negative evaluation of the self and promotes withdrawal (Tangney et al., 1996), embarrassment is typically brief and follows relatively minor social violations (Keltner & Buswell, 1997), and pride may be either authentic and achievement based or hubristic and associated with narcissism (Tracy & Robins, 2007; Weidman et al., 2016; Mercadante et al., 2021). However, it remains unclear how the properties of a socio-affective PE, including its direction, content, and whether it concerns the self or another person, shape the specific emotion that follows. Future work should test whether distinct PE profiles systematically give rise to other emotions.

Individual Differences

Due to a relatively small sample size, we only did preliminary examinations of individual differences. We found that individual participants' average PE (**Figure S3** in Online Resource 1) had substantial variability in both the magnitude and direction of these effects. This heterogeneity was particularly evident in the continuous rating data for self-videos, where some participants, on average, evaluated their own performances positively (**Figure 5**, Panel A). A similar pattern emerged in the post-task emotion endorsements: participants were nearly equally likely to characterize their own videos as embarrassing or amusing. Thus, while the overall effect indicated negative self-evaluations, there was considerable variation in how individual participants experienced the task. This heterogeneity is consistent with the embarrassment literature, which has long recognized that individuals differ in their propensity to experience embarrassment (Miller, 1995). Stable characteristics such as fear of negative evaluation, public self-consciousness, and motivation to avoid social rejection have been identified as key predictors of embarrassability (Miller, 1995). Although self-report questionnaires assessing traits and individual characteristics may help explain this variability, the present sample sizes are insufficient to support moderation analyses.

Limitations and Future Directions

Study 2 is the preferred methodological iteration, but there are still some considerations for improvement. First, although an affective measure was added in Study 2, it was limited in scope. Affect was assessed trial-by-trial along one appraisal dimension, valence, but emotional experience is widely conceptualized as multidimensional (Barrett & Russell, 1999). Additionally, the post-task emotion endorsement allowed a single emotion and intensity selection, which may obscure mixed or co-occurring emotional states and reduce reliability relative to multi-item or continuous measures (Barrett, 2006; Diamantopoulos et al., 2012). Because these emotion reports

were collected after the task rather than in the moment, they may reflect memory biases rather than real-time affective experience (Kahneman et al., 2004). Future research can add additional dimensions to the trial-by-trial assessment to increase confidence in the relationship between socio-affective PEs and embarrassment.

Second, the present findings suggest substantial heterogeneity in participants' responses to the induction, with some individuals experiencing strong negative self-related PEs while others appeared to find the experience amusing. Self-report measures of embarrassability, fear of negative evaluation, public self-consciousness, and related traits could help identify moderators of this effect and clarify whether individuals particularly sensitive to social evaluation exhibit larger or more persistent socio-affective PEs.

Third, the PGT may be well suited for use with clinical populations who have altered, delayed, or effortful processing of social and affective information (Pinkham, 2014; Senju, 2012). This vulnerability may reflect the distinctive complexity of socio-affective PEs themselves: unlike non-social errors, these signals must integrate contextual social nuances including social norms, relationship history, and the perceived expectations of interaction partners (Brown & Brüne, 2012). Consequently, populations with these impairments may exhibit attenuated or delayed socio-affective PEs, making individual differences in PE magnitude a potentially informative marker of social functioning. For example, individuals with social anxiety disorder, characterized by heightened fear of negative evaluation and rejection sensitivity, may show exaggerated and persistent PEs in response to the induction, as social anxiety is associated with negatively biased social evaluation learning and heightened self-monitoring in evaluative contexts (Miller, 1995; Rapee & Heimberg, 1997).

Fourth, the temporal dynamics of the continuous rating signal were not formally analyzed here, though this signal could capture real-time PE (**Figure 5**, Panel A). Questions remain regarding the number and timing of inflection points within each trial, the frequency and magnitude of deviations from neutrality, and the extent to which these moment-to-moment fluctuations predict overall PE scores and emotional experience. Identifying when, within a trial, critical psychological events occur will be particularly important for neuroimaging applications of this paradigm and could help establish whether changes in continuous rating trajectories constitute a valid computational marker of socio-affective PE.

Conclusion

The present studies addresses a fundamental gap in the literature: the absence of experimental paradigms designed intentionally to induce and measure socio-affective PEs as computational events in their own right. By situating embarrassment within a PE framework, PGT provides an optimized behavioral readout of socio-affective PE magnitude. While more work is needed to fully validate the paradigm and characterize the individual differences that shape these effects, the robustness of several key findings across two independent samples is encouraging. With continued development, PGT has the potential to advance our understanding of the conditions that reliably produce socio-affective PEs and the mechanisms that support real-time social error detection, with applications spanning psychological, clinical, and neuroimaging research.

Declarations

Funding: This work was externally funded.

References

- Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology*, 20(1), 1–48.
- Allman, M. J. (2011). Deficits in temporal processing associated with autistic disorder. *Frontiers in Integrative Neuroscience*, 5, 2.
- Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7), 280–289.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, 21(1), 37–46.
- Baron-Cohen, S., O’Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, 29(5), 407–418.
- Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1), 28–58.
- Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science*, 8(1), 10–14.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, 115(2), 243.
- Bhatia, S., Mellers, B., & Walasek, L. (2019). Affective responses to uncertain real-world outcomes: Sentiment change on Twitter. *PLoS ONE*, 14, e0212489.
<http://dx.doi.org/10.1371/journal.pone.0212489>

- 885 Boothby, E. J., Cooney, G., Sandstrom, G. M., & Clark, M. S. (2018). The liking gap in
 886 conversations: Do people like us more than we think? *Psychological Science*, 29(11),
 887 1742–1756.
- 888 Brown, B. R., & Garland, H. (1971). The effects of incompetency, audience acquaintanceship,
 889 and anticipated evaluative feedback on face-saving behavior. *Journal of Experimental*
 890 *Social Psychology*, 7(5), 490–502.
- 891 Brown, E. C., & Brüne, M. (2012). The role of prediction in social neuroscience. *Frontiers in*
 892 *Human Neuroscience*, 6, 147.
- 893 Crozier, W. R. (2014). Differentiating shame from embarrassment. *Emotion Review*, 6(3), 269–
 894 276.
- 895 Dahl, D. W., Manchanda, R. V., & Argo, J. J. (2001). Embarrassment in consumer purchase: The
 896 roles of social presence and purchase familiarity. *Journal of Consumer Research*, 28(3),
 897 473–481.
- 898 Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for
 899 choosing between multi-item and single-item scales for construct measurement: A
 900 predictive validity perspective. *Journal of the Academy of Marketing Science*, 40(3),
 901 434–449.
- 902 Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one’s own ignorance. In
 903 *Advances in Experimental Social Psychology* (Vol. 44, pp. 247–296). Academic Press.
- 904 Epley, N., & Dunning, D. (2000). Feeling “holier than thou”: Are self-serving assessments
 905 produced by errors in self- or social prediction? *Journal of Personality and Social*
 906 *Psychology*, 79(6), 861–875.

- 907 Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews*
 908 *Neuroscience*, 11(2), 127–138.
- 909 Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of*
 910 *Physiology-Paris*, 100(1–3), 70–87.
- 911 Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*,
 912 63(1), 287–313.
- 913 Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews*
 914 *Neuroscience*, 9(4), 304–313.
- 915 Joiner, J., Piva, M., Turrin, C., & Chang, S. W. (2017). Social learning through prediction error
 916 in the brain. *NPJ Science of Learning*, 2(1), 8.
- 917 Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey
 918 method for characterizing daily life experience: The day reconstruction method. *Science*,
 919 306(5702), 1776–1780.
- 920 Keltner, D., & Buswell, B. N. (1997). Embarrassment: Its distinct form and appeasement
 921 functions. *Psychological Bulletin*, 122(3), 250–270.
- 922 Kim, Y. H., Chiu, C. Y., & Zou, Z. (2010). Know thyself: Misperceptions of actual performance
 923 undermine achievement motivation, future performance, and subjective well-being.
 924 *Journal of Personality and Social Psychology*, 99(3), 395–409.
- 925 Krach, S., Cohrs, J. C., de Echeverría Loebell, N. C., Kircher, T., Sommer, J., Jansen, A., &
 926 Paulus, F. M. (2011). Your flaws are my pain: Linking empathy to vicarious
 927 embarrassment. *PLoS ONE*, 6(4), e18675.
- 928 Leary, M. R., Landel, J. L., & Patton, K. M. (1996). The motivated expression of embarrassment
 929 following a self-presentational predicament. *Journal of Personality*, 64(3), 619–636.

- 930 Mayer, A. V., Müller-Pinzler, L., Krach, S., & Paulus, F. M. (2020). Spinach in the teeth: How
 931 ego- and allocentric perspectives modulate neural correlates of embarrassment in the face
 932 of others' public mishaps. *Cortex*, 130, 275–289.
- 933 Mayr, E. (1974). Behavior programs and evolutionary strategies. *American Scientist*, 62(6), 650–
 934 659.
- 935 Melchers, M., Li, M., Chen, Y., Zhang, W., & Montag, C. (2015). Low empathy is associated
 936 with problematic use of the Internet: Empirical evidence from China and Germany. *Asian*
 937 *Journal of Psychiatry*, 17, 56–60.
- 938 Mercadante, E. J., & Tracy, J. L. (2024). How does it feel to be greedy? The role of pride in
 939 avaricious acquisition. *Journal of Personality*, 92(2), 565–583.
- 940 Miller, R. S. (1995). On the nature of embarrassability: Shyness, social evaluation, and social
 941 skill. *Journal of Personality*, 63(2), 315–339.
- 942 Miller, R. S. (1996). *Embarrassment: Poise and peril in everyday life*. Guilford Press.
- 943 Miller, R. S., & Leary, M. R. (1992). Social sources and interactive functions of emotion: The
 944 case of embarrassment. In M. S. Clark (Ed.), *Emotion and social behavior* (pp. 202–221).
 945 Sage.
- 946 Moll, J., Oliveira-Souza, R. D., Garrido, G. J., Bramati, I. E., Caparelli-Daquer, E. M., Paiva, M.
 947 L., ... & Grafman, J. (2007). The self as a moral agent: Linking the neural bases of social
 948 agency and moral sensitivity. *Social Neuroscience*, 2(3–4), 336–352.
- 949 Müller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frässle, S., ... & Krach, S.
 950 (2015). Neural pathways of embarrassment and their modulation by social anxiety.
 951 *NeuroImage*, 119, 252–261.

- 952 Pinkham, A. E. (2014). Social cognition in schizophrenia. *Journal of Clinical Psychiatry*,
 953 75(Suppl 2), 14–19.
- 954 Rapee, R. M., & Heimberg, R. G. (1997). A cognitive-behavioral model of anxiety in social
 955 phobia. *Behaviour Research and Therapy*, 35(8), 741–756.
- 956 Riddell, C., Nikolić, M., & Kret, M. E. (2025). Putting the social in emotions: The effect of
 957 audience presence on pride and embarrassment across ontogeny. *Developmental Science*,
 958 28(4), e70024.
- 959 Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward.
 960 *Science*, 275(5306), 1593–1599.
- 961 Sedikides, C., & Gregg, A. P. (2003). Portraits of the self. In M. A. Hogg & J. Cooper (Eds.),
 962 *Sage handbook of social psychology* (pp. 110–138). Sage.
- 963 Senju, A. (2012). Spontaneous theory of mind and its absence in autism spectrum disorders.
 964 *Neuroscientist*, 18(2), 108–113.
- 965 Shearn, D., Bergman, E., Hill, K., Abel, A., & Hinds, L. (1990). Facial coloration and
 966 temperature responses in blushing. *Psychophysiology*, 27(6), 687–693.
- 967 Shearn, D., Bergman, E., Hill, K., Abel, A., & Hinds, L. (1992). Blushing as a function of
 968 audience size. *Psychophysiology*, 29(4), 431–436.
- 969 Shepperd, J. A., & McNulty, J. K. (2002). The affective consequences of expected and
 970 unexpected outcomes. *Psychological Science*, 13, 85–88. [http://dx.doi.org/10.1111/1467-](http://dx.doi.org/10.1111/1467-9280.00416)
 971 9280.00416
- 972 Shepperd, J. A., Klein, W. M., Waters, E. A., & Weinstein, N. D. (2013). Taking stock of
 973 unrealistic optimism. *Perspectives on Psychological Science*, 8(4), 395–411.

- 974 Singh, D., & Bhushan, B. (2025). Understanding shame, guilt, embarrassment and pride: A
 975 systematic review of self-conscious emotions. *Frontiers in Psychology*, 16, 1678930.
- 976 Sturm, V. E., Ascher, E. A., Miller, B. L., & Levenson, R. W. (2008). Diminished self-conscious
 977 emotional responding in frontotemporal lobar degeneration patients. *Emotion*, 8(6), 861–
 978 869.
- 979 Tangney, J. P., & Dearing, R. L. (2002). Gender differences in morality.
- 980 Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are shame, guilt, and
 981 embarrassment distinct emotions? *Journal of Personality and Social Psychology*, 70(6),
 982 1256–1269.
- 983 Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective
 984 on mental health. *Psychological Bulletin*, 103(2), 193–210.
- 985 Tracy, J. L., & Robins, R. W. (2007). The psychological structure of pride: A tale of two facets.
 986 *Journal of Personality and Social Psychology*, 92(3), 506.
- 987 Villano, W. J., Otto, A. R., Ezie, C. E., Gillis, R., & Heller, A. S. (2020). Temporal dynamics of
 988 real-world emotion are more strongly linked to prediction error than outcome. *Journal of*
 989 *Experimental Psychology: General*, 149(9), 1755.
- 990 Weidman, A. C., Tracy, J. L., & Elliot, A. J. (2016). The benefits of following your pride:
 991 Authentic pride promotes achievement. *Journal of Personality*, 84(5), 607–62.
 992

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [02.PGTOnlineResource1.docx](#)