Title

The SocialVidStim: a video database of positive and negative social evaluation stimuli for use in social cognitive neuroscience paradigms

Permalink

https://escholarship.org/uc/item/6gj4z4wn

Journal

Social Cognitive and Affective Neuroscience, 19(1)

ISSN

1749-5016

Authors

Tully, Laura M
Blendermann, Mary
Fine, Jeffrey R
et al.

Publication Date

2024-04-12

DOI

10.1093/scan/nsae024

Peer reviewed

Social Cognitive and Affective Neuroscience, 2024, **19**(1), nsae024

DOI: https://doi.org/10.1093/scan/nsae024
Advance Access Publication Date: 10 April 2024

Original Manuscript

The SocialVidStim: a video database of positive and negative social evaluation stimuli for use in social cognitive neuroscience paradigms

Laura M. Tully,[1] Mary Blendermann,[1] Jeffrey R. Fine,[2] Lauren N. Zakskorn,[1] Matilda Fritz,[1] Gabriella E. Hamlett,[1] Shannon T. Lamb,[1] Anna K. Moody,[1] Julenne Ng,[1] Narimes Parakul,[1] Bryn M. Ritter,[1] Raisa Rahim,[1] Grace Yu,[1] and Sandra L. Taylor[2]

[1]Department of Psychiatry and Behavioral Sciences, University of California, Davis, Sacramento, CA 95817, USA
[2]Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, Sacramento, CA 95817, USA
Correspondence should be addressed to Laura M. Tully, Department of Psychiatry and Behavioral Sciences, UC Davis Imaging Research Center, 4701 X Street, Sacramento, CA 95817, USA. E-mail: tully.laura@gmail.com

Abstract

This paper describes the SocialVidStim—a database of video stimuli available to the scientific community depicting positive and negative social evaluative and neutral statements. The SocialVidStim comprises 53 diverse individuals reflecting the demographic makeup of the USA, ranging from 9 to 41 years old, saying 20–60 positive and 20–60 negative social evaluative statements (e.g. 'You are a very trustworthy/annoying person'), and 20–60 neutral statements (e.g. 'The sky is blue'), totaling 5793 videos post-production. The SocialVidStim are designed for use in behavioral and functional magetic resonance imaging paradigms, across developmental stages, in diverse populations. This study describes stimuli development and reports initial validity and reliability data on a subset videos ($N = 1890$) depicting individuals aged 18–41 years. Raters perceive videos as expected: positive videos elicit positively valenced ratings, negative videos elicit negatively valenced ratings and neutral videos are rated as neutral. Test–retest reliability data demonstrate intraclass correlations in the good-to-excellent range for negative and positive videos and the moderate range for neutral videos. We also report small effects on valence and arousal that should be considered during stimuli selection, including match between rater and actor sex and actor believability. The SocialVidStim is a resource for researchers and we offer suggestions for using the SocialVidStim in future research.

Keywords: stimulus set; social cognition; videos; fMRI; multiracial

Introduction

Interpersonal stressors such as negative social evaluation (NSE) can trigger physiological stress responses (Dickerson *et al.*, 2008), symptom exacerbations in psychotic disorders (Tully *et al.*, 2014a), and may be associated with the development of depression and anxiety (Silk *et al.*, 2012), likely due to increased sensitivity to social rejection (Silk *et al.*, 2014) and impaired cognitive control of emotion (Hooker *et al.*, 2011; Goldin *et al.*, 2014; Tully *et al.*, 2014b; Masland *et al.*, 2015; Yin *et al.*, 2015). Conversely, receiving positive social evaluation (PSE) promotes positive social interactions and prosocial behavior (Reis *et al.*, 2010; Sallquist *et al.*, 2012; Yao *et al.*, 2016); is associated with reduced hyperactivity, disruption and exclusion by peers (Sallquist *et al.*, 2012); and can induce short-term ameliorations in childhood shyness and related behaviors (Greco and Morris, 2001). Although cognitive behavioral therapy (CBT), the predominant psychosocial intervention for mood, anxiety and psychotic disorders, has been shown to improve cognitive control of emotion and reduce reactivity to negative social stimuli by targeting target frontal–limbic and frontal–parietal networks, changes in neural networks are not always associated with changes in mental health symptoms (e.g. Kumari *et al.*, 2011; Ritchey *et al.*, 2011; Yang *et al.*, 2018; Rubin-Falcone *et al.*, 2020). Understanding the neural mechanisms underlying response to negative and positive social evaluation could facilitate our understanding of vulnerability and protective factors for mental health outcomes and identify potential treatment targets.

A fundamental challenge of examining neural mechanisms underlying response to social evaluation is that the experimental environment (e.g. an MRI scanner) is inherently non-social. Consequently, researchers must rely on stimulus sets that aim to evoke neural and behavioral responses as similar as possible to responses that might occur in real-world social interactions. However, currently available and popular stimuli sets, such

Received: 6 February 2021; **Revised**: 6 January 2023; **Accepted**: 5 April 2024

as the International Affective Picture System (IAPS; Lang et al., 2005), NimStim face set (Tottenham et al., 2009), Adolphs face set (Adolphs et al., 1998) and Montreal Set of Facial Displays of Emotion (MSFDE; Beaupré and Hess, 2005) are typically static images created using predominantly anachronistic Western settings and White adults. These stimulus sets have three key drawbacks that limit their ecological validity, resulting in fMRI data that may not accurately represent how the brain responds to real-world social interactions.

First, using static stimuli might mask important findings. Behaviorally, compared to static emotional face stimuli, dynamic emotional face stimuli result in faster and more accurate emotion recognition (Calvo et al., 2016), are perceived as more intense and realistic (Zloteanu et al., 2018), and elicit stronger and more frequent emotion reactions (for reviews see: Krumhuber et al., 2013; Lander and Butcher, 2020). Dynamic stimuli also elicit greater neural activation in and stronger connectivity between social and emotional brain regions (Kilts et al., 2003; LaBar et al., 2003; Sato et al., 2004, 2012; Schultz and Pilz, 2009; Trautmann et al., 2009). Importantly, these effects may vary depending on sex, psychiatric diagnoses and neurodevelopmental status. For example, male individuals rate dynamic expressions of anger to have higher intensities than static ones (Biele and Grabowska, 2006), individuals with high social anxiety are better at recognizing negative emotions in static pictures than in dynamic animations (Torro-Alves et al., 2016), and, compared to neurotypical individuals, autistic individuals show reduced activation (Sato et al., 2004) and connectivity (Sato et al., 2012) in social brain regions (Amygdala, STS, Fusiform, mPFC) in response to dynamic *vs* static face stimuli. Given that the majority of social affective neuroscience studies use static stimuli, our current estimations of group differences in processing of social and emotional information may not be accurate, and findings may not be generalizable to the inherently dynamic setting of the real world.

Second, the majority of stimuli sets depicting people are of adults, which limits researchers' ability to investigate developmental trajectories of response to social stimuli in same-aged peers. Developmental research demonstrates age-related changes in response to peer interactions. Adolescents report increased peer socialization, sensitivity to peer influence and peer rejection compared to adults (Larson and Richards, 1991; Gardner and Steinberg, 2005; Steinberg, 2005; Choudhury et al., 2006; Pfeifer and Blakemore, 2012). Similarly, young adolescents are less successful at regulating responses to social than to nonsocial stimuli (Silvers et al., 2012), and effects of reappraisal/regulation of negative stimuli last longer as individuals age (Silvers et al., 2015), likely due to increased coupling between prefrontal regulation mechanisms and the amygdala (Silvers et al., 2017). Types of reappraisal strategies used also change over the course of development (Nook et al., 2020). Collectively, these findings demonstrate there are important age-related changes in the perception and regulation of social stimuli. However, the lack of stimuli sets that offer developmentally appropriate options for participants spanning middle childhood through adulthood limits our understanding of how these age-related changes occur and their impact on social interactions over the course of development.

Third, norming data for these stimuli sets are typically collected in relatively small samples ($N < 1000$), raising questions regarding validity and reliability of the stimuli in terms of eliciting the experimentally desired response in diverse populations. There is a critical need for new stimuli sets that are large, dynamic (e.g. videos, interaction vignettes), broadly representative of multiple cultural dimensions (race, ethnicity, age, sex assigned at birth, gender identity, sexual orientation, attractiveness, etc.), and well-validated for use across the lifespan.

To address this critical need, we created a database of video stimuli suitable for use in fMRI paradigms, across developmental stages, in diverse populations. We filmed 53 diverse individuals reflecting the demographic makeup of the USA, ranging from 9 to 41 years old, saying 20–60 positive and 20–60 negative social evaluative statements (e.g. 'You are a very trustworthy/annoying person'), and 20–60 neutral statements (e.g. 'The sky is blue'), resulting in 5793 videos post-production. To maximize usability in the fMRI environment, in which stimuli are often shown on small screens and scanner noise can interfere with audio stimuli, we created videos with high-definition audio and visual characteristics. Each video also includes ∼2 s of still, neutral expression at the start and end, allowing researchers to account for Blood Oxygen Level Dependent response to faces in the absence of social feedback. The resulting set of videos is large, and the actors depicted are diverse across multiple sociocultural dimensions, making the videos suitable for use in diverse study populations. Collection of validity and reliability data on all videos in the SocialVidStim is currently underway, with norming data for ∼33% of videos already collected and reported in this manuscript. Here we provide details regarding stimuli development and report initial validity (study 1) and reliability (study 2) data on a subset of the video stimuli ($N = 1890$) depicting individuals 18–41 years.

In this paper, we report data from two studies: study 1 (validity data) includes ratings of key video characteristics (valence, arousal, believability) on 1890 videos of actors aged 18–41 years (888 negative, 858 neutral, 144 positive) from 1781 participants, gathered via the UC Davis undergraduate study pool ($N = 1546$) and Amazon Mechanical Turk ($N = 235$). Study 2 (reliability data) includes intra-rater test–retest reliability data on a subset of these videos ($N = 226$; 84 negative, 89 neutral, 53 positive) from 390 participants, also gathered via the UC Davis undergraduate study pool. For each study, we report descriptive statistics of stimuli valence, arousal, and believability ratings, as well as analyses examining the effects of actor sex, rater sex, and participants' beliefs about themselves on the perception of social evaluation. We hypothesized that (i) participants' valence, arousal, and believability ratings would provide support for the validity of this new set of video stimuli (study 1); (ii) participants' valence and arousal ratings would be reliable across two testing sessions (study 2).

Methods

Development of the SocialVidStim

Selection of social evaluative statements

A pool of 965 statements (340 negative, 283 positive, 342 neutral) were generated by the research team for use in the SocialVidStim with adult actors. Positive and negative statements were declarative social evaluations matched for content and word length (e.g. 'Everyone likes you' *vs* 'No one likes you') with roughly equal distribution of first person (e.g, 'I would date you'/'I wouldn't date you'), second person (e.g. 'You are clever' *vs* 'You are stupid') and third person (e.g. 'Everyone thinks you are a success' *vs* 'Everyone thinks you are a failure') statements. Neutral statements were either factual statements about the world (e.g. 'The sky is blue', 'Elephants have trunks') or factual statements about the person (e.g. 'I play guitar', 'I am vegetarian'). To select the statements for filming, we collected valence ratings on all 965 statements via the UC Davis SONA Study Pool ($N = 953$ participants, 67%

female; mean age = 19.71, SD = 2.03, age range: 18–38 years) and Amazon Mechanical Turk (N = 89 participants, 54% female; mean age = 27.69, SD = 4.54, age range = 18–44 years). SONA participants received course credit for participation; MTurk participants were compensated $0.03 per statement rated. Participants were eligible for the study if they were between 18 and 45 years old and had English fluency (defined as learning English before age 5). MTurk raters also were required to live in a predominantly English-speaking country (including the USA, the UK, Ireland, New Zealand, Australia and Canada) and to have a HIT Approval Rate of 95% or higher (a measure of data quality from their prior MTurk work).

Participants rated statement valence on a 7-point Likert scale (‘How does this statement make you feel?’; 1 = very bad/upset, 4 = neither upset nor happy, 7 = very good/happy). A total of 136 negative, 116 neutral and 117 positive statements were selected based on how close they were to the desired valence: selected negative statements were the statements with average valences closest to 1; selected neutral statements had average valences closest to 4; and selected positive statements had average valences closest to 7. See Supplementary Tables 6a–c for the final list of statements filmed with adult actors.

Actors

In this paper, we present data collected on 40 adult actors (20 females; ages 18–41 years, M = 23.9, SD = 4.9). Actors were recruited via promotional flyers distributed to the UC Davis Departments of Psychology and Theatre, the Sacramento State University Department of Psychology, the Sacramento Comedy Club, and local Sacramento acting groups on Facebook. We focused our efforts on recruiting actors that represented diversity across a range of sociocultural dimensions (age, race, ethnicity, gender presentation) in order to create a diverse stimuli set. Figure 1 demonstrates the diverse demographic makeup of the SocialVidStim actors in comparison to the demographic makeup of the USA (based on 2019 data from the annual American Community Survey: https://data.census.gov/cedsci). See Supplementary materials for demographic characteristics, stimuli development details and photographs of the 40 actors included in this paper.

Video production

Actors were filmed in front of a green screen in a small, sound-controlled room with studio lighting using a Nikon D3200 with a 60 mm lens placed 74 inches in front of the actor. Adult actors were recorded saying 30–60 positive, 30–60 negative and 30–60 neutral statements, while looking directly into the camera as if talking to or looking at another person and conveying the emotion corresponding to the statement’s valence. All actors gave consent for their videos to be released for public use and were compensated for their time.

Video clips were edited using Final Cut Pro X (Apple Inc., Cupertino, CA, USA) to be ∼6 s long, starting with ∼2 s of the actor looking into the camera, then ∼2 s for statement delivery and ending with 2 s of the actor looking into the camera. Videos with poor audio or video quality, obvious mismatches in statement valence and actor delivery, or in which speech was unclear were discarded. After editing procedures, the number of SocialVidStim videos rated in this study totals 1890 (888 negative, 858 neutral, 144 positive). See Figure 2 for a depiction of workflow from stimuli creation through data collection and analysis.

Study 1: validity of the SocialVidStim

Study 1 sought to establish the validity of the SocialVidStim by collecting valence, arousal and believability ratings on 1890 (888 negative, 858 neutral, 144 positive) videos in the SocialVidStim. We aimed to collect a minimum of 30 ratings per video up to 100. After data cleaning, we report validity data on 1002 videos (429 negative, 429 neutral, 144 positive) from 1781 participants—see Figure 2 and *quality control & data exclusions* section for details on videos excluded from validity analyses.

Study 1 participants

A total of 1781 participants (60.6% female) participated in study 1: 1546 participants (62.4% female) were recruited via the UC Davis SONA Study Pool (SONA); 235 participants (47.7% female) were recruited via Amazon Mechanical Turk (MTurk). Participant sample size was driven by the goal to collect a minimum of 30 ratings per video, up to 100, with roughly equal numbers of male and female participants. SONA study pool participants received course credit for their participation; MTurk participants were compensated $0.03 cents per video rated. Participants were eligible for the study if they were between 18 and 35 years old and had English fluency (defined as learning English before age 5 years). MTurk raters also were required to live in a predominantly English-speaking country (including the USA, the UK, Ireland, New Zealand, Australia and Canada) and to have a HIT Approval Rate of ≥95% (a measure of data quality from their prior MTurk work). Detailed demographic data (age, sex, race, ethnicity, first language, country of origin and years of education) was collected for SONA participants; only age and sex were collected for MTurk participants. Demographics and sample characteristics for all participants are presented in Table 1. SONA participants were younger than MTurk participants (mean difference = 7.6; SD = 2.4). This is consistent with demographic summaries of people on MTurk indicating an average age or 37 years old (Moss *et al.*, 2023). The UC Davis Institutional Review Board approved this study. All participants provided informed consent.

Study 1 video stimuli

We collected valence, arousal and believability (i.e. quality of acting in the video) ratings for a subset of SocialVidStim videos (N = 1890; 888 negative, 858 neutral, 144 positive) from 40 different actors (20 females; ages 18–41 years, M = 23.9, SD = 4.9). Seven actors identified as Hispanic/Latinx. Actors’ race identities were as follows (note that actors could identify with more than one race): 9 Asian, 6 Black/African American, 28 Caucasian/White, 2 Hawaiian/Other Pacific Islander, 1 Native American/Alaska Native). See Supplementary Table S1 for study 1 actor demographics. Study 1 video stimuli were chosen as part of a larger effort to develop two new social cognition tasks for use in fMRI studies, including an emotion regulation/reappraisal task (Tully *et al.*, 2019) and a social reward/motivation task based on work by Crawford *et al.* (2020). As such videos that met task design requirements (e.g. equal male and female actors, demographic distribution reflecting the target population for the tasks, valence requirements) were prioritized for validity and reliability data collection. Videos presented a total of 324 unique statements: 132 negative, 117 neutral and 75 positive.

Note that 232 (86 negative, 92 neutral, 54 positive) of these videos were also rated a second time as part of study 2 (test–retest reliability); only time 1 ratings for these videos were included in study 1 analyses.

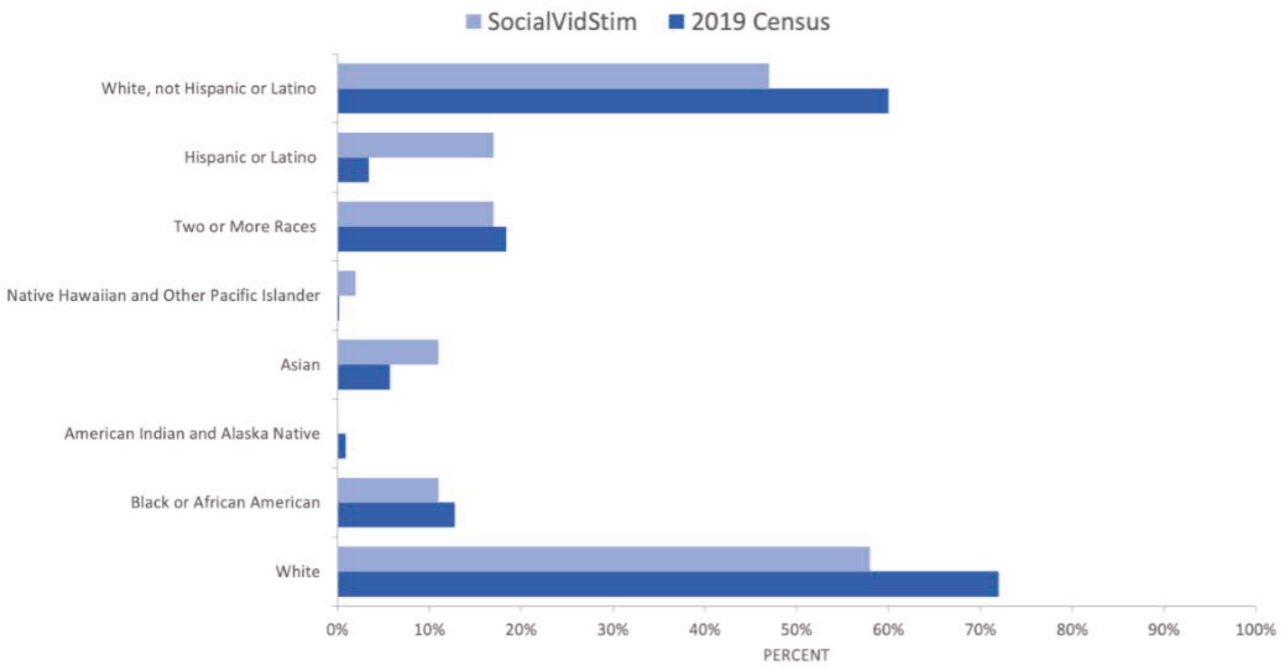

Fig. 1. Comparison of race and ethnicity characteristics of the SocialVidStim and the USA. Demographic data for the USA was obtained from the 2019 American Community Survey Demographic and Housing Data available at https://data.census.gov/cedsci.

Alt tetx: A horizontal bar chart comparing the proportion of individuals identifying as White (not Hispanic or Latino), Hispanic or Latino, two or more races, Native Hawaiian and Other Pacific Islander, Asian, American Indian and Alaska Native, Black or African American, or White (including Hispanic or Latino) in the SocialVidStim and the USA according to the USA 2019 census.

Study 1 procedure

Ratings for each video were collected online as follows: participants viewed each video on its own screen and then advanced to a separate screen to rate the video using 1–7 Likert scales on the following characteristics: valence ('How does this video make you feel?'; 1 = very bad/upset, 4 = neither upset or happy, 7 = very good/happy); arousal ('How excited or calm does this video make you feel?'; 1 = calm, completely relaxed and/or sleepy, 4 = neither calm or excited, 7 = excited, wide awake, and/or stimulated); and believability ('How believable is this video?'; 1 = extremely unbelievable, 4 = neither believable or unbelievable, 7 = extremely believable). Participants were instructed to rate believability in terms of how convincing the actor's acting of the intended valence of the video was; as such, believability can be interpreted as one proxy for acting quality. For negative and positive videos, participants also rated the extent to which they felt the statement given in the video was true of them ("How true is this of you?; 1 = definitely false, 4 = neither true or false, 7 = definitely true). We recorded the length of time participants spent on each screen (watching each video, providing ratings) for data quality control purposes (see 'Data analysis' section).

SONA participants rated videos using the online survey platform Qualtrics (Qualtrics, Provo, UT). SONA participants completed video ratings remotely online, accessing the link to the Qualtrics survey through the SONA website, except for eight surveys where data were collected during an in-person testing session on the UC Davis campus (7 test–retest surveys, 1 validity survey). Video ratings gathered via online surveys did not meaningfully differ from video ratings gathered via in-person testing sessions (average valence online = 3.7 ± 1.2; average valence in-person = 3.8 ± 1.3; average arousal online = 4.1 ± 1.0; average arousal in-person = 4.1 ± 1.1). Each Qualtrics survey contained 30–80 videos (corresponding to the number of course credits available for completing the survey) balanced for valence across sub-selections of actors. A total of 24 Qualtrics surveys were distributed: 17 surveys collected validity data only, 7 surveys collected test–retest data (time 1 ratings from these surveys were included in validity data analyses), 10 surveys included videos of all three valences and 14 surveys included negative and neutral videos only (this was done to prioritize data collection for a task development project). Video ratings gathered via surveys with negative and neutral videos only did not differ from video ratings gathered via surveys with negative, neutral and positive videos.

MTurk participants rated individual videos (i.e. a HIT) presented on the MTurk workforce recruitment platform using HyperText Markup Language (HTML) and Cascade Styling Sheets (CSS). MTurk participants only rated negative and neutral videos. MTurk participants could complete as many HITs as desired (average number of HITs = 31.6, SD = 50.5, range = 1–346).

Study 2: reliability of the SocialVidStim

Study 2 seeks to establish the test–retest reliability of the SocialVidStim. To date, test–retest valence, arousal and believability data have been collected on a subset of 226 videos (84 negative, 89 neutral, 53 positive).

Study 2 participants

A total of 390 participants (51.3% female) recruited via the UC Davis SONA study pool completed study 2. Inclusion criteria were: 18–35 years old and English fluency. Demographics and sample characteristics for study 2 participants are presented in Table 2. As in study 1, participant sample size was driven by the goal to collect a minimum of 30 ratings per video, up to 100, with roughly equal numbers of male and female participants. The UC Davis Institutional Review Board approved this study. All participants provided informed consent and received course credit for participation.

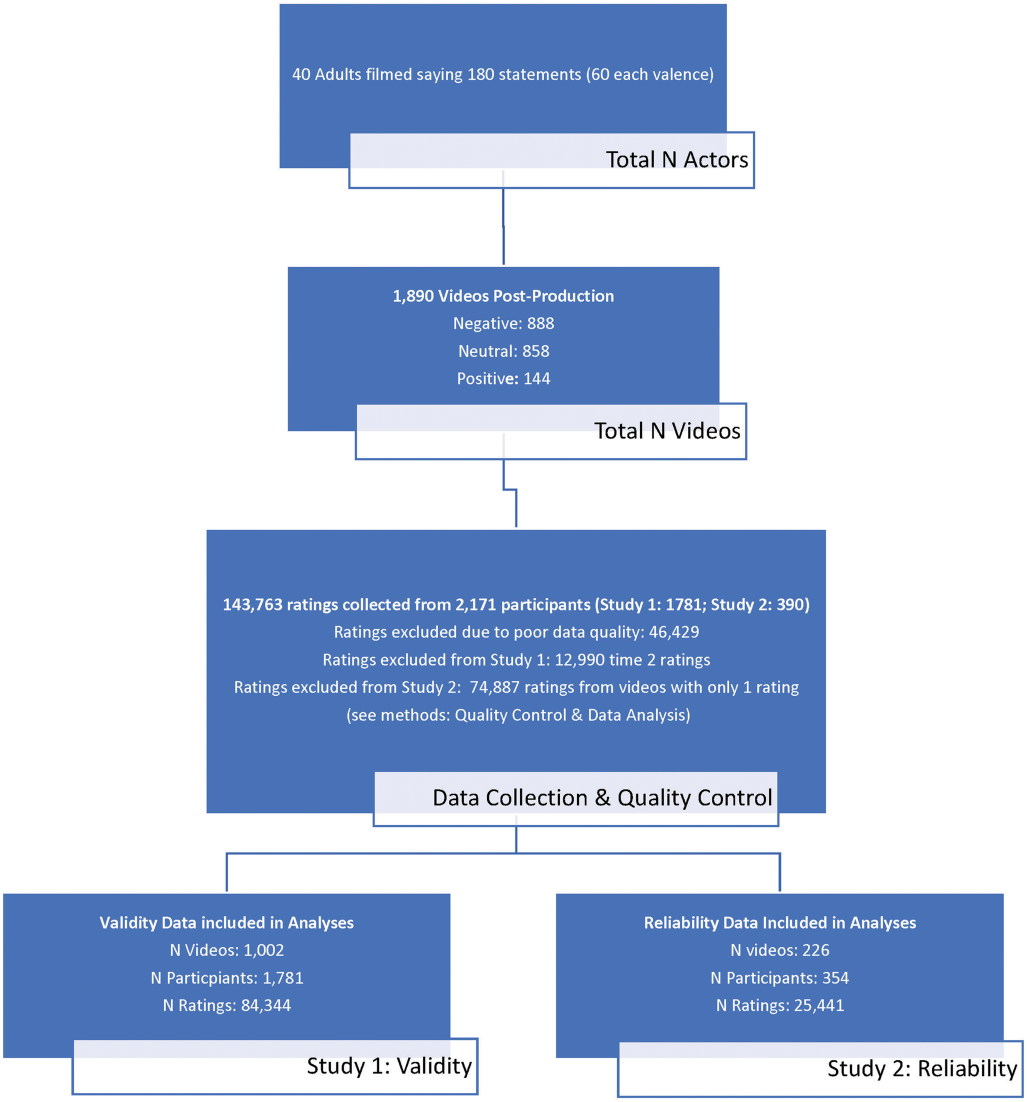
40 Adults filmed saying 180 statements (60 each valence)
Total N Actors
1,890 Videos Post-Production
Negative: 888
Neutral: 858
Positive: 144
Total N Videos
143,763 ratings collected from 2,171 participants (Study 1: 1781; Study 2: 390)
Ratings excluded due to poor data quality: 46,429
Ratings excluded from Study 1: 12,990 time 2 ratings
Ratings excluded from Study 2: 74,887 ratings from videos with only 1 rating
(see methods: Quality Control & Data Analysis)
Data Collection & Quality Control
Validity Data included in Analyses
N Videos: 1,002
N Particpiants: 1,781
N Ratings: 84,344
Study 1: Validity
Reliability Data Included in Analyses
N videos: 226
N Participants: 354
N Ratings: 25,441
Study 2: Reliability

Fig. 2. Creation and evaluation of SocialVidStim. Flow diagram depicting pathway from video creation (filming, editing), to data collection (validity, test–retest reliability), to data included in analyses (post-data cleaning). Information displayed pertains only to the 40 adult actor videos evaluated in this manuscript.

Alt tetx: A vertical flow diagram showing the pathway from video creation (filming, editing), to data collection (validity, test–retest reliability), to data included in analyses (post-data cleaning).

Study 2 video stimuli

Study 2 video stimuli included 226 videos (84 negative, 89 neutral, 53 positive) from 19 different actors (9 females; ages 18–41, M = 23.8, SD = 7.3). Three actors identified as Hispanic/Latinx. Actors' race identities were as follows (note that actors could identify with more than one race): 7 Asian, 5 Black/African American, 10 Caucasian/White, 1 Hawaiian/Other Pacific Islander, 1 Native American/Alaska Native). See Supplementary Table S2 for study 2 actor demographics. Videos presented a total of 124 unique statements: 48 negative, 42 neutral and 34 positive. Time 1 ratings for these videos were included in study 1 validity analyses.

Study 2 procedure

All test–retest reliability data were collected in-person on the UC Davis campus via group sessions of up to 20 participants at a time, supervised by research staff. Participants completed test–retest ratings via Qualtrics surveys. A total of 7 Qualtrics surveys were used to collect test–retest data; 6 of these contained 33 videos and 1 contained 34, balanced for valence across sub-selections of actors. Due to experimenter error, 2 surveys did not contain positive videos. Ratings gathered via surveys with negative and neutral videos only (average valence = 4.1 ± 1.1; average arousal = 3.6 ± 1.1) did not differ from video ratings gathered via surveys with negative, neutral and positive videos (average valence = 4.1 ± 1.1; average arousal = 3.9 ± 1.4). Each test–retest survey was divided into three parts: in part 1, participants rated each video on valence, arousal and the extent to which they felt the statement given in the video was true of them (we did not collect believability ratings in study 2 to reduce participant burden). In part 2, participants completed a series of distractor tasks for 10 min. Distractor tasks included word searches, picture searches and anagrams, and did not contain social or emotional information. In part 3, participants rated the videos presented in part 1 a second time. All videos were presented in randomized order in both parts 1 and 3. We recorded the length of time participants spent on each screen (watching each video, providing ratings) for data quality control purposes (see 'Data analysis' section).

Table 1. Demographics and sample characteristics for Study 1 (validity data)

	All participants (N =1781)		SONA (N =1546)		MTurk (N =235)	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**
N (participants)[a]	701 (39.4)	1076 (60.4)	581 (37.6)	964 (62.4)	120 (51.1)	112 (47.7)
Age (years)	21.4 ± 3.95 [18–35]	20.8 ± 3.41[18–35]	20.1 ± 2.29 [18–34]	19.9 ± 1.83 [18–35]	27.6 ± 4.37 [19–35]	28.5 ± 4.02 [19–35]
Ethnicity and race[b,c]						
Hispanic/Latinx[d]			131 (8.5)	234 (15.1)		
African American/Black			20 (1.3)	33 (2.1)		
Asian			326 (21.1)	440 (28.5)		
Hawaiian/Other Pacific Islander			16 (1.0)	26 (1.7)		
Caucasian/White			184 (11.9)	353 (22.8)		
Native American/Alaska native			8 (0.5)	12 (0.8)		
Declined to state			60 (3.9)	145 (9.4)		
First language[b,g]						
English			457 (29.6)	821 (53.1)		
Spanish			84 (5.4)	175 (11.3)		
Other			249 (16.1)	373 (24.1)		
Country of origin[e]						
USA			408 (26.4)	754 (48.8)		
Other			171 (11.1)	210 (13.6)		
Years of education[f]			13.3 ± 1.51 [11–25]	13.5 ± 1.43 [12–21]		

Race, ethnicity, language, country of origin and education data were not available for MTurk Raters. Continuous variables are presented as mean± SD [min–max]; categorical data are presented as N (% Total).;

[a] Sex data were missing for four participants.

[b] Participants could identify with multiple races and/or report multiple first languages therefore percentages may sum to greater than 100%.

[c] Race data were missing for seven participants.

[d] Ethnicity data were missing for 7 participants and 16 participants declined to state their ethnicity.;

[e] Country of origin data were missing for three participants.

[f] Education data were missing for 24 participants.

[g] Language data were missing for one participant.

Data analysis

Quality control and data exclusions.

A total of 143 763 video ratings were collected on 1890 videos (888 negative, 858 neutral, 144 positive). All ratings were examined for quality and cleaned using a standardized script in MATLAB R2019a. Data were excluded that did not meet inclusion criteria or quality control metrics: 14 260 ratings (9.92%) were excluded from raters outside the ages of 18–35 years; 3 ratings (0.002%) were excluded because the rater spent <5 s viewing the 6-s video; 3428 ratings (2.38%) were excluded because the time the rater spent on the survey ratings page was less than half the median time they spent on all the ratings pages throughout the survey; 5445 ratings (3.79%) were excluded from raters who made valence ratings that were all 4s, or all 1s and 7s across all valence types; 5279 ratings (3.67%) were excluded from raters who made arousal ratings that were all 4s, or all 1s and 7s across all valence types; 160 ratings (0.11%) were excluded from raters whose valence ratings had a standard deviation of <0.1 across all valence types; 115 ratings (0.08%) were excluded from raters whose arousal ratings had a standard deviation of <0.1 across all arousal types; 719 ratings (0.50%) were excluded because they were duplicate ratings of the same video from the same rater; and 1036 raters (0.72%) were excluded who had not provided information about their sex. For study 1 validity analyses, we excluded 12 990 time 2 ratings from the test–retest surveys (9.04%) and excluded 15 984 ratings (11.1%) because they were from videos that had fewer than 30 ratings from analysis ($n = 884$; 456 negative, 428 neutral). Thus, after quality control, a total of 84 344 ratings on 1002 videos (429 negative, 429 neutral, 144 positive) from all 1781 participants were included in validity analyses.

For study 2 reliability analyses, after applying general quality control criteria as detailed above, we excluded 74 887 video ratings from 36 raters that only had one rating (801 negative, 768 neutral, 91 positive). Thus, after quality control, a total of 25 441 ratings (13 647 time 1 and 11 794 time 2) on 226 videos (84 negative, 89 neutral, 53 positive) from 354 participants (172 males, 182 females) were included in reliability analyses.

Statistical analyses.

All statistical analyses were conducted using SAS 9.4 (SAS Institute). Quantitative variables were summarized by means ± standard deviation (SD) and categorical variables as proportions. Hypothesis tests were two-sided and evaluated at a significance level of 0.05. We analyzed the SONA and MTurk ratings together and included rate type as a factor in all models for the most powerful and efficient use of the data. For study 1, mixed-effects linear regression models were used first to determine if there was a difference in valence and arousal ratings between negative, neutral, and positive video emotion types. For these models, valence or arousal ratings were modeled as a function of video emotion type, the specific statement the rater watched nested within the video emotion type and rater type (SONA *vs* MTurk). Rater and actor were included as random effects to account for correlated responses. Second, for both valence and arousal ratings we then fitted a model that included the rater's sex and age, and the interaction between the rater's sex and the video emotion type that was watched to evaluate whether male and female individuals differentially rate the different video emotion types. A third model, for both valence and arousal ratings, included the rater's sex, actor's sex and age, and the interaction between the match between rater and actor sex and video

Table 2. Sample characteristics and demographics for study 2 (test–retest data)

	All participants (N = 354)	
	Male	Female
Total participants, *N* (% total)	172 (48.6)	182 (51.4)
Age (mean ± SD) [min—max]	19.9 ± 2.0 [18–30]	19.7 ± 1.4 [18–26]
Ethnicity and race, *N* (% total)[a]		
Hispanic/Latinx	31 (8.8)	60 (16.9)
African American/Black	1 (0.3)	4 (1.1)
Asian	111 (31.3)	82 (23.2)
Hawaiian/Other Pacific Islander	6 (1.7)	6 (1.7)
Caucasian/White	47 (13.3)	54 (15.3)
Native American/Alaska Native	3 (0.8)	1 (0.3)
Declined to State	18 (5.1)	47 (13.3)
First language, *N* (% total)[a]		
English	128 (36.2)	128 (36.2)
Spanish	23 (6.5)	48 (13.6)
Other	89 (25.1)	87 (24.6)
Country of origin, *N* (% total)		
USA	111 (31.4)	131 (37.0)
Other	61 (17.2)	51 (14.4)
Years of education (mean ± SD) [min—max]	13.3 ± 1.5 [12–19]	13.3 ± 1.3 [12–18]
Occupation, *N* (% total)		
Student	172 (48.6)	181 (51.1)
Other	0 (0.0)	1 (0.3)

All study 2 data were collected via the UC Davis SONA Study pool. Continuous variables are presented as mean± SD [min–max]; categorical data are presented as *N*(% Total).
[a]Participants could identify with multiple races and/or report multiple first languages therefore percentages may sum to >100%.

emotion type to evaluate whether the males and females watching either a male or female actor rated different video emotion types differently. We constructed contrasts to specifically evaluate sex differences in valence and arousal ratings by video emotion type following a significant interaction effect.

Mixed-effects linear regression models were also used to test whether valence and arousal ratings were associated with believability ratings and whether the participant perceived the statement as being true of them. For this analysis, we modeled valence or arousal as a function of sex, video emotion type and either believability or the extent participants rated the statement to be true of themselves, and all two- and three-way interactions among these variables. We also adjusted for rater's age, rater type and the statement that the rater watched. Rater and actor were included as random effects. For significant interaction effects, contrasts were constructed to specifically evaluate the change in believability or the extent participants rated the statement to be true of themselves, and valence and arousal ratings by sex and video emotion type.

For study 2, test–retest reliability was determined by calculating the intra-class correlations (ICC) broken down by rater sex and video emotion type. The focus of the test–retest reliability assessment was to evaluate how consistently a rater rates a given video seen on two occasions. We used the two-way mixed effect model approach to calculating the ICC where the 'judges' were the two testing occasions, and the raters were random targets. The INTRACC (1) macro was used to estimate the ICC and 95% confidence intervals.

Table 3. Average valence, arousal and believability ratings for 1002 videos of the SocialVidStim

	All raters (N = 1781)			Female raters (N = 1076)			Male raters (N = 701)		
	Negative videos	Neutral videos	Positive videos	Negative videos	Neutral videos	Positive videos	Negative videos	Neutral videos	Positive videos
Valence	2.8 ± 1.1 [1–7]	4.1 ± 0.7 [1–7]	4.8 ± 1.1 [1–7]	2.8 ± 1.1 [1–7]	4.1 ± 0.6 [1–7]	4.8 ± 1.1 [1–7]	2.9 ± 1.1 [1–7]	4.1 ± 0.7 [1–7]	4.8 ± 1.1 [1–7]
Arousal	4.4 ± 1.1 [1–7]	3.7 ± 1.0 [1–7]	4.3 ± 1.2 [1–7]	4.4 ± 1.0 [1–7]	3.8 ± 0.9 [1–7]	4.3 ± 1.1 [1–7]	4.4 ± 1.2 [1–7]	3.7 ± 1.1 [1–7]	4.2 ± 1.2 [1–7]
Believability	4.4 ± 1.6 [1–7]	4.8 ± 1.4 [1–7]	4.5 ± 1.5 [1–7]	4.4 ± 1.6 [1–7]	4.8 ± 1.4 [1–7]	4.5 ± 1.5 [1–7]	4.3 ± 1.7 [1–7]	4.9 ± 1.5 [1–7]	4.5 ± 1.5 [1–7]

84 344 ratings on 1002 videos (429 negative, 429 neutral, 144 positive) were included in validity analyses. Descriptive statistics are reported as mean ± SD [min–max].

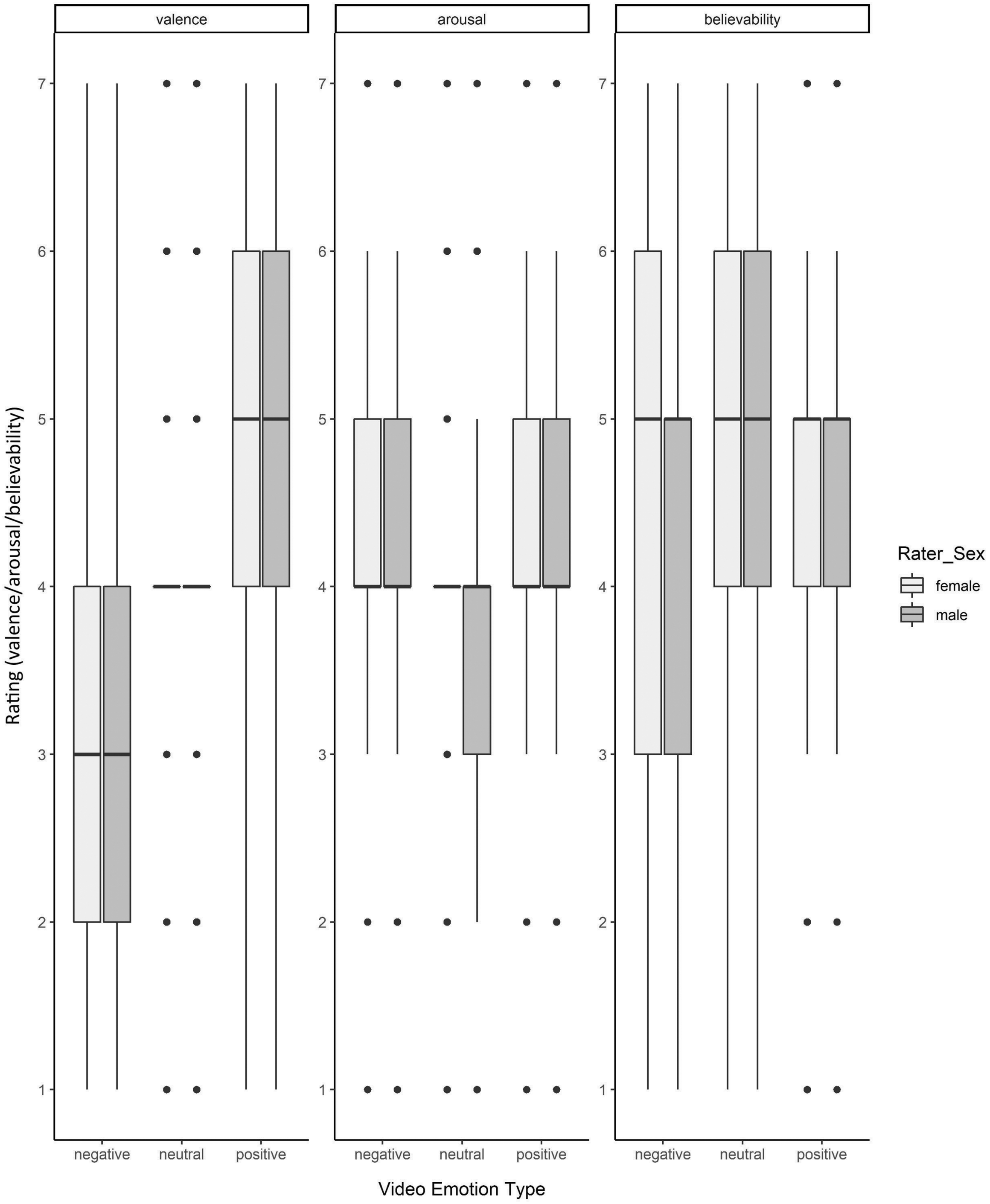

Fig. 3. Boxplots of valence, arousal, and believability ratings by video emotion type and rater sex for videos rated in Study 1. Participants rated each domain on a 1-7 Likert scale: Valence ("How does this video make you feel?": 1 = very bad/upset, 4 = neither upset or happy, 7 = very good/happy); arousal ("How excited or calm does this video make you feel?"; 1 = calm, completely relaxed and/or sleepy, 4 = neither calm or excited, 7 = excited, wide awake, and/or stimulated); believability ("How believable is this video?"; 1 = extremely unbelievable, 4 = neither believable or unbelievable, 7 = extremely believable). The boxplot includes the minimum value, 1st (lower) quartile (Q1), median, 3rd (upper) quartile (Q3), and the maximum value. Outliers are also indicated on a box plot.

Alt text: A vertical box plot comparing valence, arousal, and believability ratings by video emotion type and rate sex for videos rated in sudy 1.

Results

Study 1: validity of SocialVidStim

Effect of video emotion type on valence and arousal ratings.

Descriptive statistics for video ratings are reported in Table 3 and visualized in Figure 3. As predicted, negative videos were perceived as more negatively valenced (mean difference = –1.20; 95% CI [–1.22, –1.18]) and more arousing (mean difference = 0.68, 95% CI [0.66, 0.71]) than neutral videos. Similarly, positive videos were perceived as more positively valenced (mean difference = 0.85, 95% CI [0.82, 0.88]) and more arousing (mean difference = 0.58, 95% CI [0.54, 0.62]). Average valence rating for neutral videos was 4.1 (SD = 0.7) and average arousal rating was 3.7 (SD = 1.0) indicating that overall participants experienced neutral videos as neutral.

Examination of sex differences in valence and arousal ratings.

Analyses of sex effects revealed that, on average, males rated negative videos less negatively (mean difference = 0.06, 95% CI [0.02, 0.09]) and higher on arousal (mean difference = 0.05, 95% CI [0.01, 0.1]) compared to females. However, the magnitude of this difference (<0.1 of a point in valence/arousal) is small and may not be meaningful at the behavioral level. There were no sex differences in valence or arousal ratings for positive (valence mean difference = –0.03, 95% CI [–0.08, 0.02]; arousal mean difference = –0.03, 95% CI [–0.08, 0.02]) or neutral videos (valence mean difference = 0.04, 95% CI [–0.001, 0.08]; arousal mean difference = –0.04, 95% CI [–0.09, 0.005]).

We also tested for significant interactions between rater sex and actor sex on valence and arousal ratings. That is, we examined whether valence and/or arousal ratings varied depending on the match/mismatch between rater sex and actor sex. Results indicate some statistically significant interactions with small effect sizes, as follows.

For negative video valence ratings, female raters tended to rate female actors' videos more negatively than male raters watching videos of male actors (mean difference = –0.10, 95%CI [–0.17, –0.04]). Female raters did *not* rate female actors' negative videos more negatively than male actors' negative videos (mean difference = –0.03, 95% CI [–0.09, 0.02]) nor did male raters rate negative videos of female actors more negatively than male actors' negative videos (mean difference = –0.04, 95% CI [–0.08, 0.002]).

For negative video arousal ratings, female raters tended to rate female actors' negative videos as less arousing/exciting than male raters watching negative videos of female actors (mean difference = –0.06, 95% CI [–0.11, –0.01]). No other significant interactions between rater sex and actor sex on arousal ratings for negative videos were identified.

For positive video valence ratings, female raters tended to rate female actors' positive videos more positively than male raters watching male actors (mean difference = 0.11, 95% CI [0.04, 0.19]); female raters did not rate female actors' positive videos more positively than male actors' positive videos (mean difference = 0.05, 95% CI [–0.02, 0.12]) and there was no difference in female raters watching positive videos of female actors compared to male raters watching positive videos of female actors (mean difference = 0.00, 95% CI [–0.05, 0.05]).

For positive video arousal ratings, female raters tended to rate female actors' positive videos as more arousing/exciting compared to male raters watching videos of male actors (mean difference = 0.13, 95% CI [0.01, 0.25]); no other interactions between rater sex and actor sex on arousal ratings for positive videos were observed.

Overall, analyses indicate some statistically significant effects of sex on valence and arousal ratings, both in terms of the sex of the person watching the video and the sex of actor in the video. The pattern of results suggests that valence and arousal ratings are more affected when both the actor and rater are female compared to when the both the actor and rater are male: compared to a male pair, a female pair results in negative videos being rated more negatively and positive videos being rated more positively. However, these effects are small (~0.1 of a point in either direction).

Effect of beliefs about the self on valence and arousal ratings.

For negative and positive videos, participants rated the extent to which they believed the social evaluative statement made by the actor was true of themselves on a 1–7 Likert Scale ("How true is this of you?; 1 = definitely false, 4 = neither true or false, 7 = definitely true). Overall, participants rated negative videos as being less true of themselves (mean = 2.42, SD = 1.50, range = 1–7) and positive videos as being more true of themselves (mean = 5.00, SD = 1.35, range = 1–7). Compared to female raters, male raters tended to rate negative videos as more true of themselves (mean difference = 0.15, 95% CI [0.0.04, 0.26]), and positive videos as less true of themselves (mean differences = –0.15, 95% CI [–0.27, –0.03]), although these effect sizes are small.

To examine whether the extent to which raters believed the social evaluative statements to be true of themselves ('true-of-you' ratings) impacted valence and arousal ratings, we modeled the interaction between rater sex, true-of-you ratings and video emotion type. Our hypothesis was that the more participants believed the statement to be true of themselves, the stronger the valence/arousal ratings would be in the expected direction; that is, negative videos participants rated as more true of themselves would be rated as more negatively valenced and positive videos participants rated as more true of themselves would be rated as more positively valenced. Results only partially support this hypothesis, with small effect sizes for negative videos in the opposite direction than predicted and small effect sizes for positive videos in the predicted direction. First, we examined effects on valence ratings: for negative videos, for every 1-point *increase* in 'true-of-you' ratings, there was a 0.17 *increase* in negative video valence ratings in male (95% CI [0.15, 0.18]) and a 0.07 *increase* in female (95% CI [0.05, 0.09]) raters. That is, the more raters believed the negative statement to be true of themselves, the less negatively valenced they rated the negative statement. For positive videos, for every 1-point *increase* in 'true-of-you' ratings, there was a 0.35 *increase* in positive video valence ratings in male (95% CI [0.33, 0.37]) and a 0.46 *increase* in female (95% CI[0.44, 0.49]) raters. That is, the more the raters believed the positive statement to be true of themselves, the more positively valenced they rated the positive statement. There were small sex differences in the impact of the extent participants believed the statement to be true of themselves on valence ratings: compared to female raters, male raters rate negative videos *less* negatively (mean difference = 0.10, 95% CI [0.07, 0.12]) and positive videos *less* positively (mean difference = –0.11, 95% CI [–0.14, –0.08]) the higher they rated the statements to be true of themselves.

Next, we examined the impact of the extent participants believed the statement to be true of themselves on arousal ratings: there was no impact on arousal ratings for negative videos in male raters and a very small (but statistically significant) impact in female raters: for every 1-point *increase* in 'true-of-you' ratings, there was a 0.02 increase (95%CI [0.002, 0.03]) in their arousal ratings of negative videos. For positive videos, for every 1-point *increase* in 'true-of-you' ratings, there was a 0.11 *increase* in arousal ratings for male (95% CI [0.09, 0.13]) and a 0.20 *increase* in arousal ratings for female (95% CI [0.18, 0.23]) raters. Similar to valence ratings, compared to female raters, male raters rate positive videos lower on arousal (mean difference = –0.09, 95% CI [–0.12, –0.06]) the higher they rated the statements to be true of themselves. There were no sex differences in the impact of the extent participants believed the statement to be true of themselves on arousal ratings for negative videos.

In summary, the extent to which participants believed the statement in the videos to be true of themselves did impact valence and arousal ratings and there are some small sex differences in the extent of this impact, but effects sizes are small, particularly for negative videos, and may be negligible at the behavioral level. However, we recommend collecting data on beliefs about the self/'true of you' ratings from participants as part of experimental designs using the SocialVidStim in order to account for potential variance in participant response to stimuli.

Effect of believability of actor on valence and arousal ratings.

Believability ratings were similar across video emotion types (see Table 3), indicating how believable raters found each actor did not meaningfully vary across valences. Compared to female raters, male raters rated negative videos as less believable (mean difference = –0.26, 95% CI [–0.34, –0.18]). There were no sex differences in believability ratings for neutral or positive videos.

To examine whether perceived believability of the videos impacted valence and arousal ratings we modeled the interaction between rater sex, believability ratings, and video emotion type. Our hypothesis was that the more believable the acting in the video, the stronger the valence/arousal ratings would be in the expected direction; that is, more believable negative videos would be rated as more negatively valenced and more believable positive videos would be rated as more positively valenced. Results support this hypothesis, although effect sizes are small. First we examined believability effects on valence ratings: for negative videos, for every 1-point *increase* in believability there was a *decrease* of 0.17 in valence ratings in both male (95% CI [–0.18, –0.16]) and female (95% CI [–0.18, –0.16]) raters. For positive videos, for every 1-point *increase* in believability there was an *increase* of 0.38 in valence ratings in male (95% CI [0.37, 0.40]) and a 0.40 in female (95% CI [0.38, 0.41]) raters. We also saw small effects for neutral videos: for every 1-point *increase* in believability there was a 0.15 *increase* in valence ratings in male (95% CI [0.14, 0.16]) and a 0.11 *increase* in female (95% CI [0.10, 0.12]) raters. The impact of believability on negative and positive videos did not differ between males and females. There was a small sex difference in the impact of believability on neutral video valence ratings that is unlikely to be meaningful: for every 1-point increase in believability males rated neutral videos 0.04 higher on valence compared to females (95% CI [0.02, 0.05]).

Second, we examined believability effects on arousal ratings: results show that as believability increases arousal increases for both sexes for negative and positive videos. As with valence, effect sizes are small. For negative videos, for every 1-point *increase* in believability there was a 0.18 *increase* in arousal ratings in male (95% CI [0.17, 0.19]) and a 0.15 *increase* in female (95% CI [0.14, 0.16]) raters. For positive videos, for every 1-point *increase* in believability there was a 0.18 *increase* in arousal ratings in male (95% CI [0.16, 0.20]) and a 0.21 *increase* in female (95% CI [0.19, 0.22]) raters. For neutral videos, believability did not affect arousal ratings for female raters; male raters showed a small effect: for every 1-point *increase* in believability there was a –0.03 *decrease* (95% CI [–0.04, –0.02]) in male raters' arousal ratings for neutral videos. There were small sex differences in the impact of believability on arousal ratings that are unlikely to be meaningful: compared to females, for every 1-point *increase* in believability male participants' arousal ratings were 0.03 *higher* for negative videos (95% CI[0.01, 0.04]), 0.03 *lower* for neutral videos (95% CI[–0.05, –0.01]), and 0.03 *lower* for positive videos (95% CI[–0.05, –0.002]).

In summary, perceived believability of the videos did impact valence and arousal ratings but effect sizes are small; the largest effect size was on positive video valence ratings, in which a 1-point increase in believability was associated with less than a half-point increase in valence ratings in both male and female raters.

Table 4. Test–retest reliability of the SocialVidStim (intra-class correlations)

	Negative videos	Neutral videos	Positive videos
All raters (N = 354)	0.94 [0.92, 0.94]	0.66 [0.61, 0.71]	0.87 [0.84, 0.89]
Female raters (N = 182)	0.93 [0.91, 0.94]	0.66 [0.58, 0.72]	0.87 [0.83, 0.90]
Male raters (N = 172)	0.94 [0.92, 0.95]	0.66 [0.59, 0.73]	0.87 [0.83, 0.90]

25 441 ratings (13 647 time 1; 11 794 time 2) on 226 videos (84 negative, 89 neutral, 53 positive) from 354 participants were included in reliability analyses. Reliability was evaluated by calculating the ICC (3,1), Shrout–Fleiss reliability: fixed set. ICCs and the 95% confidence interval [lower, upper] are reported.

Study 1 results summary.

Collectively, these results demonstrate that the SocialVidStim elicit expected valence and arousal ratings from participants: negative videos are experienced as negatively valenced, positive videos are experienced as positively valenced, and neutral videos are experienced as neutral. These rating are somewhat influenced by rater sex, actor sex, the extent to which raters believe the social evaluative statement to be true of themselves and believability. Although these effects are small, we recommend that researchers consider them when selecting SocialVidStim for experimental paradigms. Descriptive statistics on valence, arousal and believability for each actor and each video are provided to researchers as part of the SocialVidStim database to help guide video selection.

Study 2: test–retest reliability results

We evaluated test–retest reliability by calculating ICCs across two ratings of each video broken down by rater sex and video emotion type. Results across all participants (N = 354) indicate good reliability for positive videos (ICC positive = 0.87, 95% CI [0.84,0.89]), excellent reliability for negative videos (ICC negative = 0.93; 95% CI [0.92,0.94]) and moderate reliability of neutral videos (ICC neutral = 0.66; 95% CI [0.61,0.71]). Test–retest reliability was similar for male and female raters. See Table 4 for all ICCs and confidence intervals.

Discussion

The SocialVidStim is a large database of video stimuli of individuals making positive and negative social evaluative statements designed for use in social cognitive neuroscience research. We explicitly designed the SocialVidStim to reflect the demographic composition of the USA so as to be suitable for use with participants from diverse backgrounds, as well as designing the stimuli to be suitable for use in fMRI as well as behavioral paradigms. In this paper, we describe the development of SocialVidStim set and reported initial validity and test–retest reliability data for a subset of SocialVidStim videos of 40 adults (N = 1890; 20 female).

Validity and reliability of the SocialVidStim

Overall results indicate that raters perceive the SocialVidStim as expected: positive videos elicit positively valenced ratings, negative videos elicit negatively valenced ratings and neutral videos are rated as neutral. We report small effects on ratings of rater sex, actor sex, video believability and the extent to which raters believe the social evaluative statement to be true of themselves. We recommend that researchers consider these factors when selecting

SocialVidStim for experimental paradigms and these data are provided as part of the stimuli set to facilitate stimuli selection. Test–retest reliability data comparing valence and arousal ratings across two testing sessions demonstrate that ICCs are in the good-to-excellent range for negative and positive videos and in the moderate range for neutral videos (Koo and Li, 2016).

We observed small sex differences in valence and arousal ratings for negative videos, where male raters tended to rate negative videos as less negative and more arousing than female raters. However, these differences are very small (an average of 0.05 of a change in valence/arousal ratings) and likely do not reflect meaningful sex differences in the perception of negative social evaluation. Results also suggest small effects of the match between rater and actor sex, particularly for females: compared to a male pair, a female pair (female actor, female rater) was associated with negative videos being rated more negatively and positive videos being rated more positively. Although prior work indicates female raters tend to have stronger and more prolonged reactivity to emotional stimuli than males (Gard and Kring, 2007), and rater responses overall are typically stronger to stimuli depicting females displaying negative emotions (Orozco and Ehlers, 1998), prior research has not examined whether this effect differs depending on match between rater sex and the sex of the person in the emotional stimuli. It should be acknowledged that these effects were small (ratings changed by ~0.1 of a point); future research is needed to determine if this finding manifests in meaningful ways at the behavioral level. We recommend researchers consider actor and participant sex when selecting SocialVidStim videos for use in experimental paradigms. We also recommend researchers include sex as a variable of interest in statistical analyses, as recommended by the National Institutes of Health (Clayton, 2018), given prior literature indicating sex differences in the perception of emotional stimuli (Orozco and Ehlers, 1998; Gard and Kring, 2007) and the well-documented sex differences in mood and anxiety disorders (Kessler *et al.*, 1993; Kornstein *et al.*, 1995; Caballo *et al.*, 2014).

Comparison to other stimuli sets

The SocialVidStim adds to a growing collection of dynamic video stimuli designed to facilitate research examining the perception and navigation of social interactions and associated social and emotional content. Examples include stimuli from The Awareness of Social Inference Test (TASIT; McDonald, 2012); the Perception of Emotions Test (POET; Kilts *et al.*, 2003)) dynamic face stimuli; and several libraries of motion-capture/point-light videos depicting human motion (e.g. Vanrie and Verfaillie, 2004; Ma *et al.*, 2006). There are also two video stimuli sets that depict social evaluative statements, similar to the SocialVidStim: the E.Vids set (Blechert *et al.*, 2015; Reichenberger *et al.*, 2015) and the social evaluation videos used by Goldin *et al.* (2014). The latter set of videos is small (48 videos) features 10 predominantly White individuals (5 male, 5 female; 7 White/Caucasian, 3 Asian), includes only negative and positive statements, and no validity/reliability data are available. The E.Vids set is a German language video set, comprising 240 videos of 10 actors each making 8 positive and 8 negative evaluative statements and 8 neutral statements. Validity data are available (Blechert *et al.*, 2015), and the stimuli have been used to evaluate the common and distinct neural mechanisms underlying processing of negative and positive social evaluation (Miedl *et al.*, 2016). The SocialVidStim builds on these existing video stimuli sets, offering the largest, most diverse set of videos available to social cognition researchers to date.

Example uses

The SocialVidStim could be used in a variety of experimental paradigms in both fMRI and behavioral settings. First, the videos are well suited for paradigms investigating reactivity to and regulation of social evaluation. We are currently investigating sex and gender differences in cognitive control of negative social evaluation in individuals with no psychiatric diagnoses and individuals with schizophrenia-spectrum diagnoses using the 'Social Evaluation Task' (SET), an emotion regulation/reappraisal paradigm adapted from Goldin *et al.* (2014). Participants view negative or neutral videos and are instructed either to immerse themselves in (reactivity) or distance themselves from (regulation) the emotional experience of receiving social evaluation and then rate how bad they feel. Preliminary results indicate sex differences in reactivity to NSE and group differences in regulation of NSE (Tully *et al.*, 2019). Paradigms examining behavioral and neural correlates during reappraisal of negative emotion are common; an advantage of the SocialVidStim is the presence of positive social evaluative statements. Future work could use the stimuli set to examine upregulation and response to positive social evaluation, which may be of value for understanding persistent negative mood in depression, negative symptoms in psychotic disorders, and elevated mood in bipolar disorders.

Relatedly, the SocialVidStim has potential for use in paradigms examining the effect of social feedback in reinforcement learning. Crawford *et al.* (2020) used the SocialVidStim to compare the effect of social, monetary, and li quid incentives on goal-directed decision making via an incentivized cued task-switching paradigm (adapted from Yee *et al.*, 2016) in a group of healthy participants. Results demonstrate that although social feedback induced changes in participants' affect, it did not induce changes in motivation/task performance, indicating a possible dissociation between affective change and motivation on a cognitive task in response to social feedback. There is a large body of literature examining the neural substrates underlying the processing of nonsocial rewards; there remain questions regarding the processing of social rewards and the impact of social evaluation on social behavior. This may be particularly important for understanding neurodiverse responses to social stimuli, as well as psychopathology characterized by social functioning difficulties. For example, future research could build on recent work examining neural response to social and nonsocial reward in individuals with schizophrenia-spectrum diagnoses (e.g. Lee *et al.*, 2019) by using the SocialVidStim as an ecologically valid stimuli set for social feedback.

Advantages of the SocialVidStim

There are several advantageous characteristics of the SocialVidStim. First, the set contains a large number of stimuli across a diverse set of actors, both in terms of race/ethnicity as well as gender and personality expression. Second, because the set of videos evaluated in this paper includes actors aged 18 through 41 years these videos are suitable for use in studies with adolescents/young adults and adults, and therefore can be used in both longitudinal and cross-sectional studies examining developmental effects from late adolescence onward. Third, both negative and positive social evaluation stimuli are included in the set along with neutral statements for comparison conditions to facilitate the disentangling of mechanisms underlying perception and effects of negative *vs* positive social stimuli. Fourth, the videos are suitable for use in both behavioral and fMRI paradigms; stimuli design considerations included high-definition video and audio to maximize stimuli quality in the MRI environment.

Limitations

There are some limitations of the SocialVidStim. First, the videos are suitable for use in English-speaking populations only. Similarly, it is unclear how sociocultural differences between English-speaking communities (e.g. Australian *vs* British *vs* American *vs* Canadian) impact perception of the stimuli, since the vast majority of our data collection was conducted with individuals living in the USA. Further information is needed regarding how different cultures perceive different types of social evaluation. One recommendation is that researchers outside the USA who are interested in using the SocialVidStim collect valence and arousal data on the videos they select for use to account for possible sociocultural differences. A second shortcoming is that not all actors were perceived as equally believable, and our results demonstrate that believability affected valence and arousal ratings in the expected direction: more believable negative videos were perceived as more negatively valenced and more arousing; more believable positive videos were perceived as more positively valenced and more arousing. Effect sizes are small-to-medium, ranging from a 0.17 to 0.40 change in valence ratings and from 0.15 to 0.21 change in arousal ratings, thus believability should be considered as part of video stimuli selection. We report believability statistics for each actor in the database to enable researchers to choose videos that are most likely to have the intended effect in the experimental paradigm, both in terms of valence/arousal and in terms of actor quality/believability.

Data reported here also have limitations. First, we only report validity and reliability data for a subset of the SocialVidStim, and only on videos of adult actors. Future efforts include continued validity data collection on the remaining ∼3000 videos, including specific recruitment of raters aged 9–17 years old to provide validity and reliability data on videos with our youth actors. All SocialVidStim videos with ratings are available to researchers on request from the authors.

Summary and future directions

We hope the SocialVidStim can be a resource for social cognitive neuroscientists seeking to examine neural and behavioral mechanisms underlying perception of social evaluation.

Supplementary data

Supplementary data is available at *SCAN* online.

Data availability

Researchers interested in the SocialVidStim can either contact the authors or visit https://peplab.ucdavis.edu.

Funding

This work was supported by a UC Davis Behavioral Health Center of Excellence (BHCOE) Early Career Award awarded to L.M.T. and funded by CA Prop. 63 (the Mental Health Services Act) and the BHCOE, and a Building Interdisciplinary Research Careers in Women's Health (BIRCWH) award (K12 HD051958) awarded to L.M.T., funded by the National Institute of Child Health and Human Development (NICHD), Office of Research on Women's Health, Office of Dietary Supplements, and the National Institute of Aging. In addition, this work is supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1 TR0001860.

Conflict of interest

The authors declared that they had no conflict of interest with respect to their authorship or the publication of this article.

Acknowledgements

We thank the actors and video raters for their contribution to science. We thank Dennis Thompson, Scott Martin, and Chantel Oah for their assistance with database management.

References

Adolphs, R., Tranel, D., Damasio, A.R. (1998). The human amygdala in social judgment. *Nature*, **393**(6684), 470–4.

Beaupré, M.G., Hess, U. (2005). Cross-cultural emotion recognition among Canadian ethnic groups. *Journal of Cross-Cultural Psychology*, **36**(3), 355–70.

Biele, C., Grabowska, A. (2006). Sex differences in perception of emotion intensity in dynamic and static facial expressions. *Experimental Brain Research*, **171**(1), 1–6.

Blechert, J., Schwitalla, M., Wilhelm, F.H. (2015). Ein Video-Set zur experimentellen Untersuchung von Emotionen bei sozialen Interaktionen: Validierung und erste Daten zu neuronalen Effekten. *Zeitschrift Für Psychiatrie, Psychologie Und Psychotherapie*, **61**(2), 81–91.

Caballo, V.E., Salazar, I.C., Irurtia, M.J., Arias, B., Hofmann, S.G., Team, C.-A.R. (2014). Differences in social anxiety between men and women across 18 countries. *Personality and Individual Differences*, **64**, 35–40.

Calvo, M.G., Avero, P., Fernández-Martín, A., Recio, G. (2016). Recognition thresholds for static and dynamic emotional faces. *Emotion*, **16**(8), 1186–200.

Choudhury, S., Blakemore, S.-J., Charman, T. (2006). Social cognitive development during adolescence. *Social Cognitive & Affective Neuroscience*, **1**(3), 165–74.

Clayton, J.A. (2018). Applying the new SABV (sex as a biological variable) policy to research and clinical care. *Physiology and Behavior*, **187**, 2–5.

Crawford, J.L., Yee, D.M., Hallenbeck, H.W., *et al.* (2020). Dissociable effects of monetary, liquid, and social incentives on motivation and cognitive control. *Frontiers in Psychology*, **11**, 2212.

Dickerson, S.S., Mycek, P.J., Zaldivar, F. (2008). Negative social evaluation, but not mere social presence, elicits cortisol responses to a laboratory stressor task. *Health Psychology*, **27**(1), 116–21.

Gard, M.G., Kring, A.M. (2007). Sex differences in the time course of emotion. *Emotion*, **7**(2), 429–37.

Gardner, M., Steinberg, L. (2005). Peer influence on risk taking, risk preference, and risky decision making in adolescence and adulthood: an experimental study. *Developmental Psychology*, **41**(4), 625–35.

Goldin, P.R., Ziv, M., Jazaieri, H., Weeks, J., Heimberg, R.G., Gross, J.J. (2014). Impact of cognitive-behavioral therapy for social anxiety disorder on the neural bases of emotional reactivity to and regulation of social evaluation. *Behaviour Research and Therapy*, **62**, 97–106.

Greco, L.A., Morris, T.L. (2001). Treating childhood shyness and related behavior: empirically evaluated approaches to promote positive social interactions. *Clinical Child and Family Psychology Review*, **4**(4), 299–318.

Hooker, C.I., Tully, L.M., Verosky, S.C., Fisher, M., Holland, C., Vinogradov, S. (2011). Can I trust you? Negative affective priming influences social judgments in schizophrenia. *Journal of Abnormal Psychology*, **120**(1), 98–107.

Kessler, R.C., McGonagle, K.A., Swartz, M., Blazer, D.G., Nelson, C.B. (1993). Sex and depression in the National Comorbidity Survey I: lifetime prevalence, chronicity and recurrence. *Journal of Affective Disorders*, **29**(2-3), 85–96.

Kilts, C.D., Egan, G., Gideon, D.A., Ely, T.D., Hoffman, J.M. (2003). Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions. *Neuroimage*, **18**(1), 156–68.

Koo, T.K., Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, **15**(2), 155–63.

Kornstein, S.G., Schatzberg, A.F., Yonkers, K.A., *et al*. (1995). Gender differences in presentation of chronic major depression. *Psychopharmacology Bulletin*, **31**(4), 711–8.

Krumhuber, E.G. Kappas, A., Manstead, A.S. (2013). Effects of dynamic aspects of facial expressions: a review. *Emotion Review*, **5**(1), 41–6.

Kumari, V., Fannon, D., Peters, E.R., *et al*. (2011). Neural changes following cognitive behaviour therapy for psychosis: a longitudinal study. *Brain*, **134**(8), 2396–407.

LaBar, K.S., Crupain, M.J., Voyvodic, J.T., McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cerebral Cortex*, **13**(10), 1023–33.

Lander, K., Butcher, N. (2020). Recognising genuine from posed facial expressions: exploring the role of dynamic information and face familiarity. *Frontiers in Psychology*, **11**, 532794.

Lang, P.J., Bradley, M.M., Cuthbert, B.N. (2005). International affective picture system (IAPS): affective ratings of pictures and instruction manual. *Technical Report A-6*. Gainesville, FL.

Larson, R., Richards, M.H. (1991). Daily companionship in late childhood and early adolescence: changing developmental contexts. *Child Development*, **62**(2), 284–300.

Lee, J., Jimenez, A.M., Reavis, E.A., Horan, W.P., Wynn, J.K., Green, M.F. (2019). Reduced neural sensitivity to social vs nonsocial reward in schizophrenia. *Schizophrenia Bulletin*, **45**(3), 620–8.

Ma, Y., Paterson, H.M., Pollick, F.E. (2006). A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior Research Methods*, **38**(1), 134–41.

Masland, S.R., Hooley, J.M., Tully, L.M., Dearing, K., Gotlib, I.H. (2015). Cognitive-processing biases in individuals high on perceived criticism. *Clinical Psychological Science*, **3**(1), 3–14.

McDonald, S. (2012). New frontiers in neuropsychological assessment: assessing social perception using a standardised instrument, The Awareness of Social Inference Test. *Australian Psychologist*, **47**(1), 39–48.

Miedl, S.F., Blechert, J., Klackl, J., *et al*. (2016). Criticism hurts everybody, praise only some: common and specific neural responses to approving and disapproving social-evaluative videos. *Neuroimage*, **132**, 138–47.

Moss, A.J., Rosenzweig, C., Robinson, J., Jaffe, S.N., Litman, L. (2023). Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages. *Behavior Research Methods*, **55**(8), 4048–67.

Nook, E.C., Vidal Bustamante, C.M., Cho, H.Y., Somerville, L.H. (2020). Use of linguistic distancing and cognitive reappraisal strategies during emotion regulation in children, adolescents, and young adults. *Emotion*, **20**(4), 525–40.

Orozco, S., Ehlers, C.L. (1998). Gender differences in electrophysiological responses to facial stimuli. *Biological Psychiatry*, **44**(4), 281–9.

Pfeifer, J.H., Blakemore, S.-J. (2012). Adolescent social cognitive and affective neuroscience: past, present, and future. *Social Cognitive & Affective Neuroscience*, **7**(1), 1–10.

Reichenberger, J., Wiggert, N., Wilhelm, F.H., Weeks, J.W., Blechert, J. (2015). "Don't put me down but don't be too nice to me either": fear of positive vs. negative evaluation and responses to positive vs. negative social-evaluative films. *Journal of Behavior Therapy and Experimental Psychiatry*, **46**, 164–9.

Reis, H.T., Smith, S.M., Carmichael, C.L., *et al*. (2010). Are you happy for me? How sharing positive events with others provides personal and interpersonal benefits. *Journal of Personality and Social Psychology*, **99**(2), 311–29.

Ritchey, M., Dolcos, F., Eddington, K.M., Strauman, T.J., Cabeza, R. (2011). Neural correlates of emotional processing in depression: changes with cognitive behavioral therapy and predictors of treatment response. *Journal of Psychiatric Research*, **45**(5), 577–87.

Rubin-Falcone, H., Weber, J., Kishon, R., *et al*. (2020). Neural predictors and effects of cognitive behavioral therapy for depression: the role of emotional reactivity and regulation. *Psychological Medicine*, **50**(1), 146–60.

Sallquist, J., DiDonato, M.D., Hanish, L.D., Martin, C.L., Fabes, R.A. (2012). The importance of mutual positive expressivity in social adjustment: understanding the role of peers and gender. *Emotion*, **12**(2), 304–13.

Sato, W., Kochiyama, T., Yoshikawa, S., Naito, E., Matsumura, M. (2004). Enhanced neural activity in response to dynamic facial expressions of emotion: an fMRI study. *Cognitive Brain Research*, **20**(1), 81–91.

Sato, W., Toichi, M., Uono, S., Kochiyama, T. (2012). Impaired social brain network for processing dynamic facial expressions in autism spectrum disorders. *BMC Neuroscience*, **13**(1), 99.

Schultz, J., Pilz, K.S. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research*, **194**(3), 465–75.

Silk, J.S., Davis, S., McMakin, D.L., Dahl, R.E., Forbes, E.E. (2012). Why do anxious children become depressed teenagers?: The role of social evaluative threat and reward processing. *Psychological Medicine*, **42**(10), 2095–107.

Silk, J.S., Siegle, G.J., Lee, K.H., Nelson, E.E., Stroud, L.R., Dahl, R.E. (2014). Increased neural response to peer rejection associated with adolescent depression and pubertal development. *Social Cognitive & Affective Neuroscience*, **9**(11), 1798–807.

Silvers, J.A., Insel, C., Powers, A., *et al*. (2017). The transition from childhood to adolescence is marked by a general decrease in amygdala reactivity and an affect-specific ventral-to-dorsal shift in medial prefrontal recruitment. *Developmental Cognitive Neuroscience*, **25**, 128–37.

Silvers, J.A., McRae, K., Gabrieli, J.D., Gross, J.J., Remy, K.A., Ochsner, K.N. (2012). Age-related differences in emotional reactivity, regulation, and rejection sensitivity in adolescence. *Emotion*, **12**(6), 1235–47.

Silvers, J.A., Shu, J., Hubbard, A.D., Weber, J., Ochsner, K.N. (2015). Concurrent and lasting effects of emotion regulation on amygdala response in adolescence and young adulthood. *Developmental Science*, **18**(5), 771–84.

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences*, **9**(2), 69–74.

Torro-Alves, N., Bezerra, I.A.D.O., Rodrigues, M.R., Machado-de-sousa, J.P., Osório, F.D.L., Crippa, J.A. (2016). Facial emotion recognition in social anxiety: the influence of dynamic information. *Psychology & Neuroscience*, **9**(1), 1.

Tottenham, N., Tanaka, J., Leon, A.C., McCarry, T., Nurse, M., Hare, T.A. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research*, **168**(3), 242–9.

Trautmann, S.A., Fehr, T., Herrmann, M. (2009). Emotions in motion: dynamic compared to static facial expressions of disgust and happiness reveal more widespread emotion-specific activations. *Brain Research*, **1284**, 100–15.

Tully, L.M., Iosif, A.-M., Blendermann, M., Chahal, R., Guyer, A.E., Carter, C.S. (2019). Sex differences in cognitive control of emotion and their contribution to sex differences in schizophrenia symptom profiles. In: Paper presented at the Building Interdisciplinary Careers in Women's Health, NIH Office of Research in Women's Health Annual Meeting, NIH, Bethesda MD.

Tully, L.M., Lincoln, S.H., Hooker, C.I. (2014a). Lateral prefrontal cortex activity during cognitive control of emotion predicts response to social stress in schizophrenia. *NeuroImage: Clinical*, **6**, 43–53.

Tully, L.M., Lincoln, S.H., Liyanage-Don, N., Hooker, C.I. (2014b). Impaired cognitive control mediates the relationship between cortical thickness of the superior frontal gyrus and role functioning in schizophrenia. *Schizophrenia Research*, **152**(2), 358–64.

Vanrie, J., Verfaillie, K. (2004). Perception of biological motion: a stimulus set of human point-light actions. *Behavior Research Methods, Instruments, & Computers*, **36**(4), 625–9.

Yang, Z., Oathes, D.J., Linn, K.A., *et al*. (2018). Cognitive behavioral therapy is associated with enhanced cognitive control network activity in major depression and posttraumatic stress disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, **3**(4), 311–9.

Yao, Z., Yu, R., Du, W.-B. (2016). The spreading of social energy: how exposure to positive and negative social news affects behavior. *PLoS One*, **11**(6), e0156062.

Yee, D.M., Krug, M.K., Allen, A.Z., Braver, T.S. (2016). Humans integrate monetary and liquid incentives to motivate cognitive task performance. *Frontiers in Psychology*, **6**, 2037.

Yin, H., Tully, L.M., Lincoln, S.H., Hooker, C.I. (2015). Adults with high social anhedonia have altered neural connectivity with ventral lateral prefrontal cortex when processing positive social signals. *Frontiers in Human Neuroscience*, **9**, 469.

Zloteanu, M., Krumhuber, E.G. Richardson, D.C. (2018). Detecting genuine and deliberate displays of surprise in static and dynamic faces. *Frontiers in Psychology*, **9**(1184), 366823.

Social Cognitive and Affective Neuroscience, 2024, **19**(1), nsae024
DOI: https://doi.org/10.1093/scan/nsae024
Advance Access Publication Date: 10 April 2024
Original Manuscript
Received: 6 February 2021; **Revised**: 6 January 2023; **Accepted**: 5 April 2024