STEMorph: A Set of Morphed Emotional Face Stimuli from Angry to Happy Derived from *NimStim*

Mohammad Ebrahim Katebi[1,2*], Mohammad Hossein Ghafari[3*], and Tara Ghafari[4,5]

[1] School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran

[2] School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

[3] School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[4] Department of Experimental Psychology, University of Oxford

[5] Centre for Integrative Neuroscience, Oxford Centre for Human Brain Activity, Department of Psychiatry, University of Oxford, UK

Author Note

[*] These authors contributed equally.

Correspondence concerning this article should be addressed to Tara Ghafari, Department of Experimental Psychology, University of Oxford, Oxford, UK, OX1 2JD. Email: tara.ghafari@psych.ox.ac.uk

Abstract

Emotion recognition through facial expressions is crucial for interpreting social cues. However, it is often influenced by biases, i.e., systematic recognition advantages for particular emotions. Nevertheless, these biases are inconsistently reported across studies, likely due to methodological variations which underline the necessity for a standardized approach. Traditional face morphing methods can create unnatural-looking stimuli, and may confound the interpretation of emotions. To address this issue, we here introduce *STEMorph*, a validated stimulus set based on the NimStim set. We employed neutral-anchored morphing and neural-network-generated masks to ensure the natural appearance and integrity of the depicted emotions. We validated our stimulus set by having participants rate each face on a 9-point scale ranging from angry to happy, assessing the perceived emotional intensity. *STEMorph's* validity was confirmed through linear regression analysis, showing a strong correlation between subjective ratings and targeted emotional states. Moreover, aligning with previous research highlighting gender as a key factor in emotion recognition, *STEMorph* also exhibits variations in ratings across gender subgroups. *STEMorph's* reliability was confirmed through a two-week follow-up rating session with a subgroup of the same participants. By introducing a controlled, validated, and ecologically valid stimulus set of emotional faces, *STEMorph* paves the way for further investigations into facial emotion recognition, deepening our understanding of this vital aspect of human interaction.

Keywords: *STEMorph*, Stimulus Set, Facial Emotional Expressions, Morphed Stimuli, *NimStim*

Introduction

Emotion recognition is a fundamental aspect of human social interaction, allowing individuals to interpret and respond appropriately to the emotional states of others. Facial expressions serve as a primary modality for conveying emotions, providing valuable cues about an individual's internal states (1). However, despite its significance, the process of emotion recognition is not devoid of biases. These biases manifest as systematic recognition advantages for particular emotions and have significant implications for both clinical and non-clinical populations (2).

Recognition biases can vary across different conditions. Individuals with major depressive disorder (MDD) tend to interpret neutral or ambiguous faces as sad or threatening, reflecting a negativity bias. This aligns with findings that people with MDD are more attuned to negative emotions and less responsive to positive ones, which impairs their ability to interpret emotional cues (3, 4). Such biases may reinforce mood-congruent cognitive distortions, contributing to the maintenance and exacerbation of depressive symptoms (5, 6). Similarly, individuals with generalized anxiety disorder (GAD) and social anxiety disorder (SAD) exhibit hypervigilance toward negative emotional cues, such as anger or fear, which can lead to misinterpretations of facial expressions and reinforce maladaptive emotional responses (7, 8). Therefore, these biases can significantly impact social interactions and clinical outcomes; emphasizing the importance of their precise measurement for both research and clinical applications.

Studies often report inconsistent results regarding the direction and magnitude of emotional recognition biases (9). For instance, one study found that healthy individuals did not demonstrate a bias toward happiness or sadness when presented with neutral facial expressions (10). While another study showed that healthy participants tended to perceive neutral faces as expressing happiness, whereas depressed individuals interpreted them as neutral (8). It is uncertain whether these mixed findings reflect genuine individual differences in emotion perception or are due to methodological variations between the studies (11). Particularly, the quality and ecological validity of facial stimuli can

significantly impact participants' responses. While standardized photographic stimuli are commonly used for their control over visual properties, they may not fully capture the complexities of real-world facial expressions, which can lead to discrepancies in findings. Research indicates that more naturalistic stimuli, such as dynamic or spontaneous facial expressions, enhance ecological validity but introduce variability that complicates experimental control (12, 13). Consequently, a key challenge lies in creating stimuli that maintain natural appearance while allowing precise control over emotional expressions (12, 13).

One widely used approach to generating a continuum of emotions is through face morphing techniques. Face morphing is an image processing technique employed to achieve transformation from one face to another and facilitates a gradual transition between two facial images. By morphing facial features between two extreme emotional expressions, researchers generate a spectrum of face images with ambiguous emotional expressions for participants to interpret (14, 15). However, traditional face morphing presents challenges, particularly concerning the formation of unrealistic and unnatural facial stimuli (15) which may confound interpretations and obscure real emotion recognition processes.

To maintain naturalism, proper masking of the face images (extracting the face area while removing the background) is required, as morphing introduces artifacts in non-face regions. Using no mask (full head) or simple oval/square masks which are employed by many studies (12) may cause distortions that risk emotional content loss, since various facial features hold different degrees of significance depending on the type of expression (16, 17). For instance, the mouth region is particularly salient for recognizing happiness, while the eyes are more critical for identifying fear or sadness (18). These regional differences can lead to biases in emotion perception, as individuals may focus on different facial features based on cultural or individual preferences (17). These underscore the necessity for stimuli preserving anatomical fidelity through methods like neutral-anchored morphing (11) and standardized masking. However, there are no existing publicly available stimulus

set without these methodological confounds (12, 15, 19). While sets like the Karolinska Directed Emotional Faces (KDEF) (20) and the Pictures of Facial Affect (POFA) (21) provide valuable resources, they often rely on basic morphing techniques that can produce artifacts, particularly around facial features crucial for emotion recognition such as the eyes and mouth. Furthermore, many existing sets use simple geometric masks (oval or rectangular) that can create artificial boundaries and reduce ecological validity (12, 15). To address this gap, we developed and validated a new stimulus set utilizing neutral-anchored morphing alongside neural-network-generated masks, to optimize naturalism and coherence.

We derived the stimuli from the *NimStim set of facial expressions* (22) which is a comprehensive collection of facial images portraying a spectrum of emotions including happy, angry, and neutral expressions. The selection of anger and happiness as the target emotions for this study was based on their distinct valence and arousal profiles, which make them suitable for exploring biases in emotion recognition. Additionally, these emotions are commonly used in emotion research (2, 11), facilitating comparisons with existing literature. Investigating the transition between these emotions allows us to examine biases in emotional perception across a clear positive-negative emotional gradient, which is particularly relevant for studies on mood disorders and social cognition (2). We adopt a nuanced approach to morphing emotional faces by generating distinct naturalistic face masks for each image using neural networks. Furthermore, we utilize the neutral facial expression as a central anchor point to enhance the naturalistic quality of the stimuli. We then evaluate the validity and reliability of *STEMorph* by asking a group of participants to provide ratings of the emotional content of these stimuli. In doing so, the present paper has a dual but unified purpose: to introduce *STEMorph* as a publicly available stimulus set designed for research on facial emotion recognition, and to report the psychophysical validation that establishes its perceptual coherence and ecological fidelity.

Importantly, prior literature indicates that both the gender of the face and the gender of the perceiver influence emotion recognition. Female faces have been consistently rated as more emotionally expressive and displaying greater happiness than comparable male faces (17, 23). Simultaneously, female perceivers tend to demonstrate greater sensitivity to emotional gradations in facial stimuli, yielding steeper psychometric slopes when rating morphed expressions (24). Including both face gender and participant gender in our validation framework therefore serves two purposes: it enables convergent validity testing by checking whether *STEMorph* reproduces established gender-related effects, and it provides normative reference data to facilitate comparisons across studies using different participant demographics.

Method

Participants

A total of 50 participants (age range: 18–26 years; mean age: 21.5; 22 females) were recruited for the validity testing. All participants were medical students from either Tehran University of Medical Sciences (TUMS) or Shahid Beheshti University of Medical Sciences (SBUMS). All participants reported having normal or corrected-to-normal vision and no history of diagnosed neurological or psychiatric disorders. Written informed consent was obtained from each participant prior to participation, and all were compensated financially for their time. The study received ethical approval from the Ethics Committee of Shahid Beheshti University of Medical Sciences.

Experiment design & Face Generation

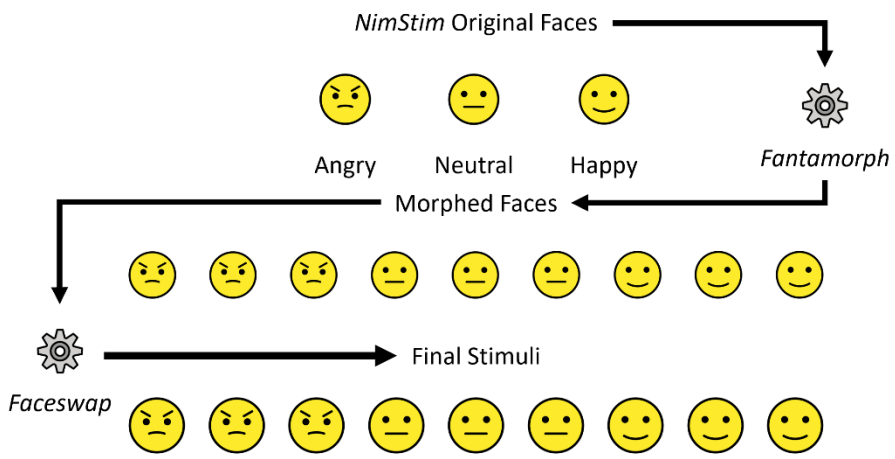
We used face morphing to produce a linear transition of emotions from anger to happiness using neutral expression as an intermediary point. The stimuli were derived from the *NimStim set of facial expressions (22)*. Each emotional expression has an open-mouth and closed-mouth variation in *NimStim*, from which we selected closed-mouth variations. *Abrosoft Fantamorph (25)* software was used for morphing, involving nine steps to transition from the angry face to neutrality and subsequently to the happy face. We used the original facial expressions of anger, happiness, and

neutral as reference points and generated three intermediate emotional states between each original face.

We selected 22 individuals (11 males and 11 females) from *NimStim*, prioritizing those whose facial expressions facilitated smooth and natural morphing while avoiding individuals with extreme facial expressions (highly exaggerated or intense displays of emotions). The final set is composed of 198 distinct faces, representing 9 different emotional steps for each of the 22 individuals (Figure 1).

Figure 1

Example Stimuli from the Set.



Note. Top row illustrates the two ends of the spectrum (happy and angry) along with the middle step (neutral). Middle row shows the intermediary morphing stages without masking. Bottom row shows the final outcome with corresponding neural-network-generated masks applied to each morphing step.

To address artifacts introduced in the ears, hair, and other non-landmark regions of the face during conventional face morphing, we generated distinct naturalistic face masks for each image using the *FaceSwap (26)* software, which utilizes the *Bisenet-FP (27)* neural network. *BiSeNet-FP* is a modified lightweight neural network designed for face parsing. The model was set to not include hair or ears and extract only the face region. Manual editing was performed to correct for any artifacts

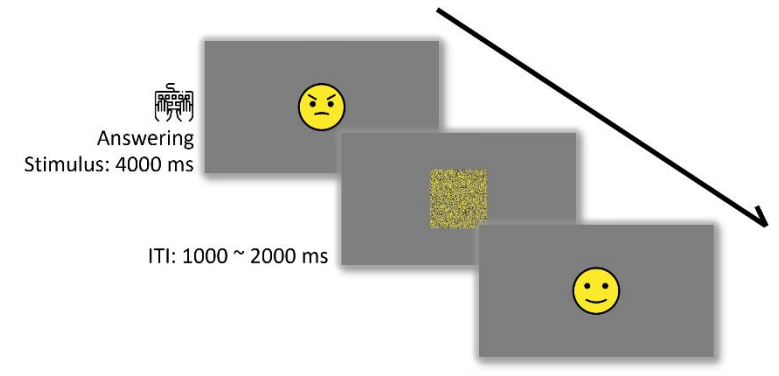
arising from stray hair (e.g., a strand of hair overlapping the face resulted in a split in the face) or other extraneous features.

Experimental Setup

Participants were asked to rate the emotional state of each presented stimulus on a 9-point scale from angry to happy. The task was developed and implemented using *Psychtoolbox* (28) within *MATLAB R2022a* (Figure 2). Participants were comfortably seated in a dimly lit room facing a laptop monitor. The resolution of monitor screen was 1920 × 1080 and it was positioned ~54 cm from the participant. Participants were instructed to freely observe the images without fixation requirements.

Figure 2

Schematic presentation of the experimental design.



Note. Stimuli appeared at the center of the screen, and participants rated the emotional intensity of each face within a 4000 ms viewing window without fixation restrictions. Each stimulus was followed by a randomized inter-trial interval (ITI) of 1000 to 2000 ms, during which a noise mask was displayed.

The task started with pressing a key, after which participants were asked to rate the emotional state of each presented face based on a 9-point scale where 1 indicated the angriest emotional state and 9 indicated the happiest emotional state. The exact instructions were: 'Please rate the emotional state of each face on a scale from 1 to 9, where 1 represents the angriest expression and 9 represents the happiest expression.' To control for motor and numerical biases, the scale was reversed for half of the participants. For these individuals, the instructions were adjusted accordingly, and their responses

were rescaled during analysis to align with the original scale (1 = angry, 9 = happy). Each participant had a 4-second window to provide their rating; in the absence of a response within this timeframe, the task automatically proceeded to the subsequent face.

Each of the 198 stimuli was presented twice in a randomized order, resulting in a total of 396 trials per participant, interspersed by two 2-minute breaks. Following each trial, a randomized inter-trial interval (ITI) of 1000 to 2000 ms ensued, during which a noise mask was displayed at the center of the screen to prevent from interference between trials and minimize priming effects (29, 30). To generate the noise mask the stimulus presented in the previous trial was scrambled into 4-pixel squares.

Reliability Assessment

To assess reliability, the same behavioral task was readministered to a subgroup of 23 participants two weeks following their initial participation. This subgroup consisted of participants who agreed to return for retesting (age range: 18-26, mean age: 22, 5 females).

Data Analysis

Validity

To investigate the relationship between morph levels and subjective ratings while accounting for gender effects, we conducted an Ordinary Least Squares (OLS) linear regression analysis using *statsmodels (31)*. The model included morph step (1-9), face gender (male/female), participant gender (male/female), and all two-way interactions as predictors of participants' emotional ratings as below:

Equation (1):

$$\boldsymbol{Participans' \ Rating} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{MS} + \boldsymbol{\beta}_2\boldsymbol{FG} + \boldsymbol{\beta}_3\boldsymbol{PG} + \boldsymbol{\beta}_4(\boldsymbol{MS} \times \boldsymbol{FG}) + \boldsymbol{\beta}_5(\boldsymbol{MS} \times \boldsymbol{PG}) + \boldsymbol{\beta}_6(\boldsymbol{FG} \times \boldsymbol{PG}) + \boldsymbol{\varepsilon}$$

where MS corresponds to Morph Step (1-9), FG to Face Gender (0=female, 1=male), PG to Participant Gender (0=female, 1=male), $\boldsymbol{\beta}$ to coefficients of each regressor, and ε to error term. Given that this is an effort to validate a stimulus set, all collected participants were included in analyses, with

no outlier removal being performed. We then calculated partial residuals for morph steps by removing variance explained by gender-related terms to isolate the morph-rating relationship.

To quantify evidence strength against null models, we computed Bayes Factors using the *BayesFactor* (32) R package. Statistical power (1- β error) was determined through *G*Power* (33) with $\alpha = 0.05$ relative to a non-correlation model ($\rho\ H_0 = 0$). To visualize partial regression plots we generated violin plots using matplotlib (34) and seaborn (35).

Gender Subgroup Analysis

To examine gender-specific effects on emotion perception, we extended our primary regression framework (Equation 1) through partial regression analysis of participant gender and face gender subgroups. We computed condition-specific partial residuals from the full model (Equation 1) to isolate morph-rating relationships within gender subgroups.

For participant gender subgroups, we calculated partial residuals by retaining morph step, participant gender (Female/Male), and their interaction (Morph Step × Participant Gender), while regressing out variance associated with face gender effects (Face Gender + Morph Step × Face Gender + Face Gender × Participant Gender) (Equation 2). Additionally, in face gender subgroup analysis, morph step, face gender (Female/Male), and their interaction (Morph Step × Face Gender) were preserved while variance attributable to participant gender effects (Participant Gender + Morph Step × Participant Gender + Face Gender × Participant Gender) were removed (Equation 3). This approach enabled examination of gender-specific emotion perception while controlling for cross-gender interactions.

Equation (2)

$$\boldsymbol{Participant\ Gender\ Partial\ \ Rating = Particiapans'\ Rating - (\beta_2 FG\ +\ \beta_4(MS \times FG)\ +\ \beta_6(FG \times PG))}$$

Equation (3)

$$\boldsymbol{Face\ Gender\ Partial\ Rating = Particiapans'\ Rating - (\beta_3 PG\ +\ \beta_5(MS \times PG) + \beta_6(FG \times PG))}$$

Where MS corresponds to Morph Step (1-9), FG to Face Gender (0=female, 1=male), PG to Participant Gender (0=female, 1=male), $\boldsymbol{\beta}$ *to coefficients of each regressor and ε to error term.*

Reliability

To investigate reliability, we analyzed data from participants who engaged in the retesting (n = 23) and compared it with their validity data. We exactly matched trials from the two tasks based on the presented face and performed OLS linear regression between participants' ratings in the validity task and the retest conducted two weeks later as below:

Equation (4):

$$\boldsymbol{Particiapans' \ Retest \ Rating} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{VR} + \boldsymbol{\varepsilon}$$

where VR corresponds to Validity Ratings (1-9), $\boldsymbol{\beta}$ to coefficient of VR, and ε to error term. We then calculated the Bayes Factor and Statistical power using the methods previously described.

Results

The aim of this study was to develop and validate a new stimulus set for emotion recognition utilizing neutral-anchored face morphing and neural-network-generated masks to optimize naturalism and coherence. A total of 50 participants were recruited to rate the emotional content of the face stimuli on a 9-point scale ranging from angry to happy with a subset of 23 participants returning for a reliability assessment two weeks later.

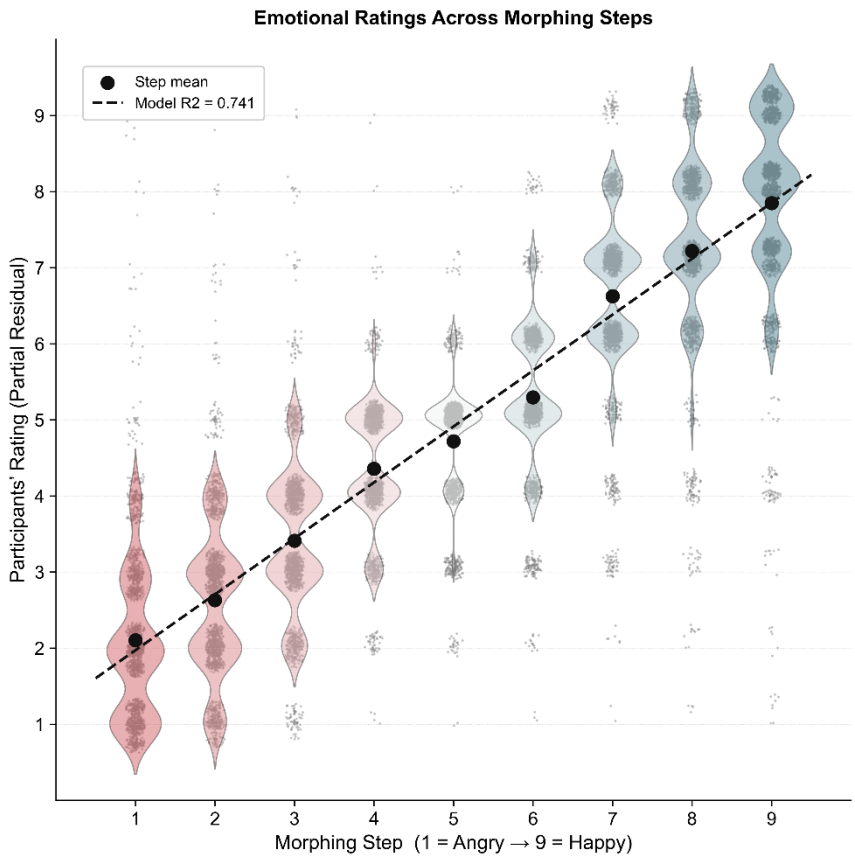
Validity

We employed linear regression to examine the emotional ratings provided by participants (mean = 4.85, SD = 2.14) against the emotional morph level of the stimulus set, while accounting for face gender, participant gender, and their interactions (Table 1). The model revealed a strong positive association between morph step and emotional ratings, with a coefficient estimate of 0.735 (95% CI [0.724, 0.746], $p < 0.001$). The model explained a substantial proportion of variance in ratings ($R^2 = 0.741$, $p < 0.001$; Figure 3). The Bayes factor comparing the model to the null hypothesis ($R^2 = 0$)

approached infinity, indicating strong evidence in favor of the model. Furthermore, the statistical power (1 − β error) relative to a non-correlation model (ρ H_0 = 0) was 0.999, demonstrating robust sensitivity to detect the observed effect.

Figure 3

Validity of emotional ratings across morphing steps.



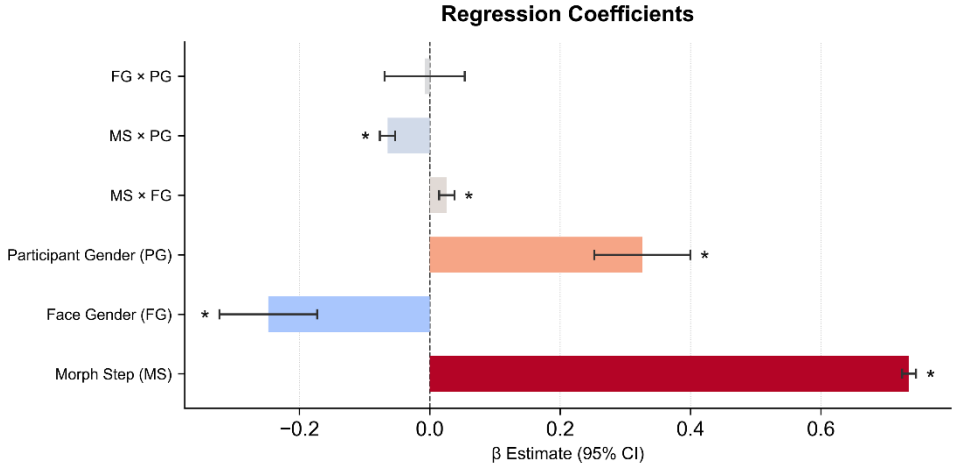
.

Note. Morphing steps range from angry (1) to happy (9) on the x-axis. The y-axis presents participants' emotional ratings as partial residuals after removing variance explained by gender-related terms (face gender, participant gender, and interactions). Violin plots show kernel density estimates of rating distributions at each morph step. Extreme outliers (±2SD from morph-level means) were excluded only for visualization clarity purposes (all data retained in statistical models). Small grey dots depict raw residuals, large black dots indicate mean ratings, and the dashed line shows the regression slope. The equation reflecting the full mode is: Participants' Rating = 0.735 × MS + -0.248 × FG + 0.326 × PG + 0.026 × (MS×FG) + -0.065 × (MS×PG) + -0.008 ×

(FG×PG) + 1.236 where MS corresponds to Morph Step (1-9), FG to Face Gender (0=female, 1=male) and PG to Participant Gender (0=female, 1=male). Colors correspond to morph steps.

Figure 4

Regression coefficients for the model.



.

Note. Beta estimates (95% CIs) from the OLS regression model predicting emotional ratings. Predictors include Morph Step (MS; 1–9), Face Gender (FG; 0 = female, 1 = male), Participant Gender (PG; 0 = female, 1 = male), and their two-way interactions. Asterisks denote terms significant at α = 0.05.

Table 1

Ordinary Least Squares Regression Results.

Predictor	*β*	*SE*	*t*	*P > |t|*	*[0.025*	*0.975]*
const	1.236	0.032	38.715	**<0.001**	1.174	1.299
Morph Step	0.735	0.005	134.882	**<0.001**	0.724	0.746
Face_Gender	-0.248	0.038	-6.485	**<0.001**	-0.323	-0.173
Participant_Gender	0.326	0.038	8.665	**<0.001**	0.252	0.400
MS_FG	0.026	0.006	4.324	**<0.001**	0.014	0.038
MS_PG	-0.065	0.006	-10.669	**<0.001**	-0.077	-0.053
FG_PG	-0.008	0.031	-0.252	0.801	-0.069	0.054

Note. OLS regression results from the full interactive model (Equation 1). The model explains 74.1% of variance in emotional ratings ($R^2 = 0.741$, N = 19,596 observations). Morph Step (MS) represents emotion intensity progression (1: angry to 9: happy). Face/Participant Gender coded 0 = Female, 1 = Male. Significant terms ($\alpha = 0.05$) marked with asterisks. The non-significant Face × Participant Gender interaction ($p = 0.801$) indicates additive rather than multiplicative gender effects.

Gender Subgroup Validity Analysis

To assess the robustness of validity across gender dimensions, we included gender of face stimuli and gender of participant in our model as described above (Equation 1). Significant effects were observed for face gender ($\beta = -0.248$, $p < 0.001$), participant gender ($\beta = 0.326$, $p < 0.001$), and the interactions between morph step and face gender ($\beta = 0.026$, $p < 0.001$) as well as morph step and participant gender ($\beta = -0.065$, $p < 0.001$; see Figure 4). The interaction between face gender and participant gender was not significant ($\beta = -0.008$, $p = 0.801$). In other words, female faces received higher ratings than male faces, whereas male participants provided higher ratings than females.

To evaluate the validity of participants' ratings in gender subgroups, we analyzed partial residual for each subgroup. Strong predictive validity across subgroups was observed. Morph step slopes for face gender subgroups (Figure 5a) were nearly identical (R^2 female faces: 0.759; R^2 male faces: 0.759), with slopes and interactions as below:

Female Faces: 1.24 + 0.74 × Morph Step; Male Faces: 0.99 + 0.76 × Morph Step

Participant gender subgroups exhibited greater divergence where female raters showed steeper slopes ($R^2 = 0.748$) than male raters ($R^2 = 0.721$, Figure 5b), suggesting gender-related differences in emotional rating:

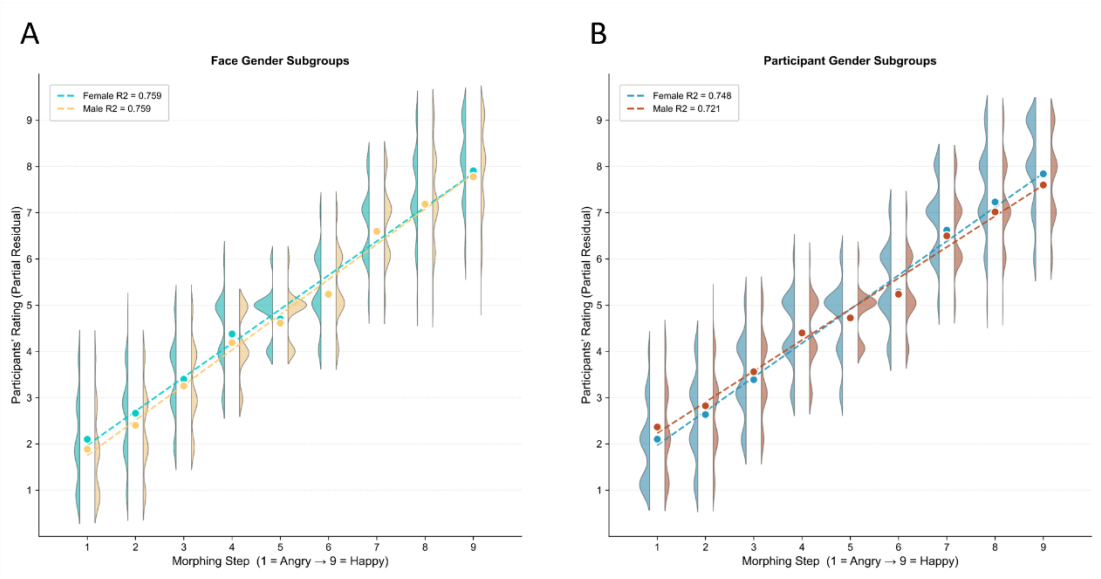
Female Participants: 1.24 + 0.74 × Morph Step; Male Participants: 1.56 + 0.67 × Morph Step

The strong R^2 of the partial regression analysis across subgroups confirm robust validity across gender dimensions while highlighting nuanced differences in how participant gender relates to emotion perception accuracy.

To evaluate whether the face gender effect reflects a broad, identity-consistent phenomenon or is driven by specific model faces, mean emotional ratings were calculated for each of the 22 *NimStim* identities by averaging across all nine morph steps and all participants (see Supplementary Figure 1a). Female identities produced a mean of 4.92 (SD = 0.16; range: 4.62–5.14), while male identities produced a mean of 4.79 (SD = 0.12; range: 4.59–4.92). Of the 11 female identities, 8 (73%) were rated above the male mean. A one-tailed Mann–Whitney U test confirmed a significant directional effect, U = 87.5, p = 0.041 (Supplementary Figure 1b). These results suggest the gender difference is distributed relatively broadly across identities rather than being attributable to a small number of cases, though the overlapping within-gender ranges indicate that identity-level characteristics also contribute to variability in perceived emotional intensity.

Figure 5

Gender-based subgroup analysis of the validity data.



Note. A) Face gender subgroups: Female faces (blue) and male faces (orange) show comparable morph-rating slopes derived from partial regression analysis. B) Participant gender subgroups: Female participants (blue) demonstrate steeper slope than male participants (orange) in partial residuals. Violin plots display distributions of gender-specific partial residuals at each morph step (1: angry to 9: happy), with dashed lines showing ordinary least squares fits to residualized data. Large markers indicate mean ratings per morph step.

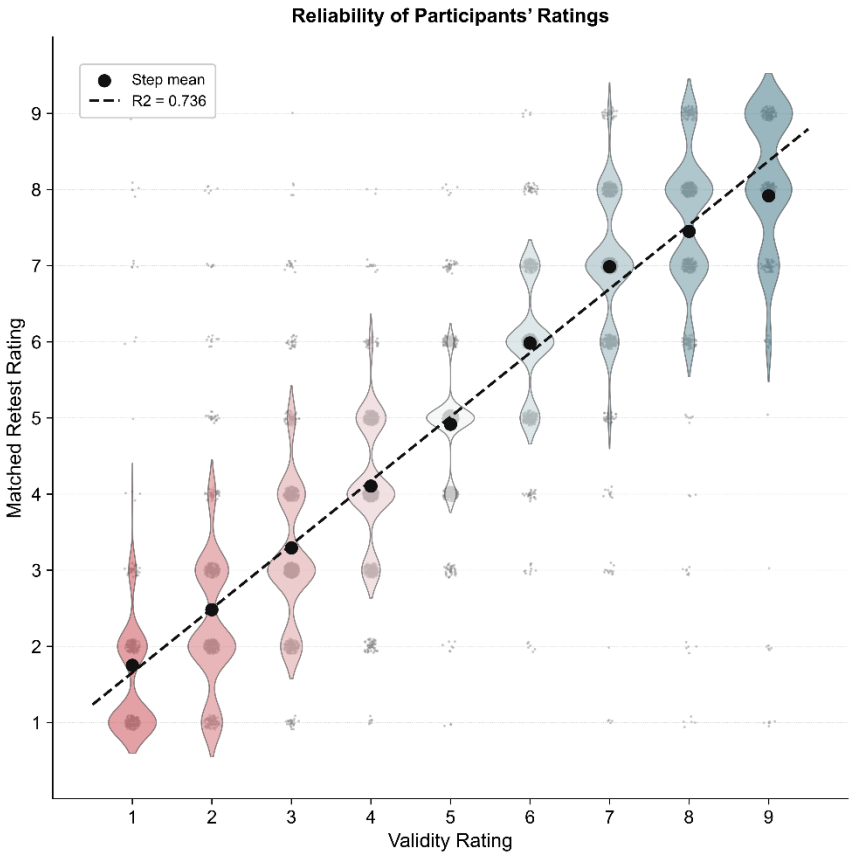
Extreme outliers (±2SD from morph-level means) were excluded only for visualization clarity purposes (all data retained in statistical models). Trendline equations reflect slope estimates from residualized relationships, with shared R^2 value from the full interactive model.

Reliability

To evaluate reliability, we examined the relationship between the validity and the retest ratings using a linear regression model. This model had a coefficient of determination (R^2) of 0.736 (p-value < 0.001) (Figure 6) and the coefficient estimate for the validity answers was 0.84, with an intercept of 0.814. The high R^2 value indicates strong test-retest reliability, suggesting minimal variability in interpretations of the stimulus set across multiple exposures. The Bayes factor, comparing the calculated regression to the null hypothesis ($R^2 = 0$), was infinite, showing strong evidence in favor of H1. Moreover, the calculated statistical power (1 – β error) in comparison with the non-correlation model (Correlation ρ H0 = 0) was 0.999, demonstrating robust sensitivity to detect the observed consistency in participants' ratings.

Figure 6

Reliability of participants’ ratings between the initial test and the retest session.



Note. The x-axis denotes the ratings from the first task (validity ratings), and the y-axis corresponds to ratings from the retest. The violin plots depict the distribution and probability density of ratings at each step, with individual trials displayed by small grey dots. The means of the retest ratings are marked with black dots, and the regression line is presented by a dashed line. The regression outcome and R^2 values are displayed in the figure. The colors are used for visualization purposes only.

Discussion

In this paper, we introduced *STEMorph*, a newly developed stimulus set based on the *NimStim* dataset, featuring facial expressions transitioning from anger to happiness. We employed neutral-anchored morphing and neural-network-generated masking and evaluated the validity and reliability of this dataset.

Participants' subjective ratings showed a robust positive correlation with the intended emotional states, indicating that our morphing methodology successfully created a naturalistic continuum of emotional transformations. This finding aligns with prior research emphasizing the

importance of realistic stimuli in emotion recognition studies (15). Moreover, *STEMorph* addresses limitations in earlier datasets, such as artifacts caused by basic morphing techniques, thereby enhancing ecological validity.

Furthermore, *STEMorph's* ability to capture gender differences in emotion recognition provides a valuable tool for addressing gaps in gender-specific emotion perception research. These results align with previous studies suggesting that facial emotion recognition is influenced by the gender of the stimuli (23) and the perceiver (24, 36, 37). Prior research has consistently reported gender-based variations in emotion recognition (17) highlighting the importance of using stimuli that account for such differences. Although the direction of the face gender effect is consistent with prior literature, per-identity normative analysis reveals meaningful within-gender variability in mean ratings (see Supplementary Figure 1), indicating that individual model characteristics contribute to the pattern alongside the gender-level effect. Researchers using *STEMorph* are therefore encouraged to consult the per-identity normative values when selecting or counterbalancing stimuli. Additionally, the high test-retest reliability observed in our study suggests minimal variability in interpretations across multiple exposures. This finding underscores the importance of well-designed stimuli in yielding consistent results in emotion perception research.

Our methodological choices also reflect advancements over traditional approaches. Specifically, utilizing a neutral expression as an intermediary anchor point minimized bias at the central position and facilitated smoother transitions between angry and happy expressions. This approach aligns with findings from earlier studies exploring emotional gradients (2) and builds on work demonstrating that neutral anchoring enhances naturalism in morphed stimuli (11). Perceptually, direct angry-to-happy morphs without a natural midpoint anchor tend to produce artefactual images at the transitional centre, where opposing muscle-group activations — corrugator contraction associated with anger and zygomaticus major activation associated with happiness — must coexist within a single face. The result can appear uncanny rather than genuinely ambiguous,

introducing a confound that is independent of the intended emotional content. This concern is supported by (11), who empirically demonstrated that the morphing strategy systematically influences emotion recognition bias, in part because perceptually incoherent midpoints distort the psychometric function. Conceptually, the neutral expression provides a semantically meaningful zero-point on the valence dimension, which is particularly valuable in bias research where the interest is the direction and magnitude of displacement relative to an emotional baseline.

Additionally, by generating distinct neural-network-based masks for each image, we mitigated issues like distortions in non-facial regions (e.g., hair or ears), which are common in datasets relying on simpler masking techniques (16). Moreover, while removing the background may reduce the real-world context of the stimuli, this step eliminates distractions, focusing solely on facial features for emotion recognition. These innovations ensure that *STEMorph* preserves anatomical fidelity while maintaining focus on facial features critical for emotion recognition.

Importantly, by making the *STEMorph* stimulus set publicly available along with all associated data and codes, we facilitate broader adoption and encourage further research in this critical domain of human interaction. This transparency aligns with calls within the literature for accessible resources to advance emotion recognition studies (9). Furthermore, *STEMorph's* design principles could inform future efforts to create ecologically valid stimuli for diverse applications, including clinical diagnostics and machine learning algorithms for facial expression analysis.

Despite these strengths, this study has limitations that warrant consideration. While our strong linear relationships provide a promising foundation for future research, further examination across diverse samples is essential to assess the generalizability and external validity of our findings. For instance, expanding participant demographics beyond young adults of similar ethnic backgrounds could reveal how age or cultural differences influence emotion perception biases. Additionally, restricting participant evaluations to a scale from anger to happiness may simplify responses. Nonetheless, we believe that focusing on the anger-happiness continuum provides a foundation for

future studies exploring other emotional transitions with more flexible response formats. Emotions such as sadness or fear may exhibit unique perceptual biases requiring tailored stimulus sets. Therefore, developing complementary transitions between various emotions could enrich our understanding of emotion recognition processes across different contexts. Furthermore, while the neutral-anchored morphing approach has clear perceptual advantages, it is worth acknowledging that anchoring to a neutral midpoint may reduce the degree of ambiguity at the central morph steps relative to a direct angry-to-happy continuum. Additionally, the current sample comprised exclusively Iranian medical students, which may limit cross-cultural and cross-demographic generalizability of the normative data.

In conclusion, our results underscore the importance of methodological rigor in stimulus design within the field of emotion recognition. The incorporation of neutral-anchored morphing and neural-network-generated masks effectively addresses limitations observed in earlier studies regarding the unnatural appearance of stimuli. By enhancing the naturalism and coherence of facial stimuli, our work significantly contributes to the ongoing efforts to improve methodologies in studying facial expressions. Furthermore, variations in emotional ratings across gender subgroups highlight the relevance of considering both participant and facial expressor characteristics in emotion recognition research. We believe that *STEMorph* can serve as a valuable resource for future studies exploring the dynamics of emotion perception across diverse populations.

Declarations

Funding

This study was supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z and 203139/A/16/Z).

Conflicts of interest

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval

Approval was obtained from the ethics committee of Shahid Beheshti University of Medical Sciences (IR.SBMU.MSP.REC.1396.757). The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate

Informed written consent was obtained from all individual participants included in the study. Facial stimuli were reused from the *NimStim* Set of Facial Expressions (22), in which participants provided informed consent for the publication of their facial images, covering their reuse.

Data and Code availability

The data acquired and codes for the psychophysics task and data analysis performed in this study can be found in the following GitHub repository: https://github.com/Mekatebi/STEMorph

To use the stimulus set, visit the *NimStim Set of Facial Expressions* website and follow the instructions to request access: https://danlab.psychology.columbia.edu/content/nimstim-set-facial-expressions

The stimulus set is also available from the corresponding author on reasonable request.

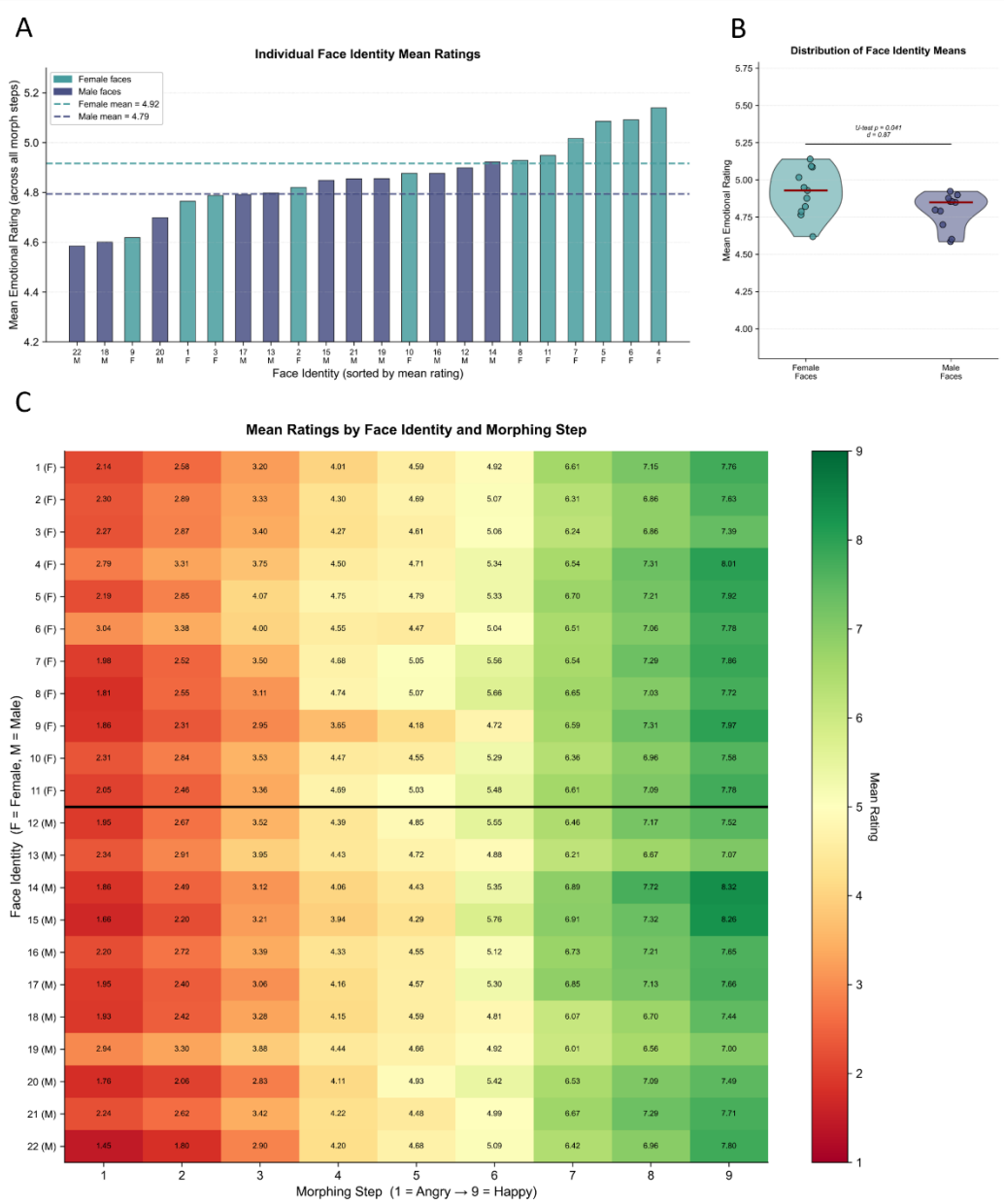
Author Contributions

MEK and MHG performed research, MEK analyzed data, all designed research and wrote the paper.

Supplementary Materials

Supplementary Figure 1

Face Identity Rating Profiles.



Note. A) Mean emotional ratings averaged across all nine morph steps for each of the 22 face identities, sorted in ascending order. Female faces are shown in teal and male faces in navy. Dashed horizontal lines indicate the overall mean for female and male faces separately. Identity labels on the x-axis report the person number and gender abbreviation (F = Female, M = Male). B) Distribution of mean emotional ratings for female (teal) and male (navy) face identities. Violin shape represents the kernel density estimate. Individual face identity means are overlaid as jittered points. The red horizontal line within each violin marks the group median. Statistical comparisons: Mann–Whitney U test p-value (one-tailed, female > male), and Cohen's d effect size. C) Mean

emotional ratings for each of the 22 NimStim face identities (rows) across the nine morph steps (columns). Female faces (F, persons 1–11) are shown in the upper section and male faces (M, persons 12–22) in the lower section. Cell color reflects mean rating magnitude on a red–yellow–green scale; cell values are rounded to two decimal places.

References

1. Russell JA. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull. 1994;115(1):102-41.

2. Silvia PJ, Allan WD, Beauchamp DL, Maschauer EL, Workman JO. Biased Recognition of Happy Facial Expressions in Social Anxiety. Journal of Social and Clinical Psychology. 2006;25(6):585-602.

3. Penton-Voak IS, Munafò MR, Looi CY. Biased facial-emotion perception in mental health disorders: A possible target for psychological intervention? Current Directions in Psychological Science. 2017;26(3):294-301.

4. Leppänen JM. Emotional information processing in mood disorders: a review of behavioral and neuroimaging findings. Curr Opin Psychiatry. 2006;19(1):34-9.

5. Harmer CJ, Goodwin GM, Cowen PJ. Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. British Journal of Psychiatry. 2009;195(2):102-8.

6. Penton-Voak IS, Bate H, Lewis G, Munafò MR. Effects of emotion perception training on mood in undergraduate students: randomised controlled trial. Br J Psychiatry. 2012;201(1):71-2.

7. Waters AM, Mogg K, Bradley BP, Pine DS. Attentional bias for emotional faces in children with generalized anxiety disorder. J Am Acad Child Adolesc Psychiatry. 2008;47(4):435-42.

8. Surguladze SA, Young AW, Senior C, Brebion G, Travis MJ, Phillips ML. Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. Neuropsychology. 2004;18(2):212-8.

9. Münkler P, Rothkirch M, Dalati Y, Schmack K, Sterzer P. Biased Recognition of Facial Affect in Patients with Major Depressive Disorder Reflects Clinical State. PLOS ONE. 2015;10(6):e0129863.

10. Leppänen J, Milders M, Bell J, Terriere E, Hietanen J. Depression biases the recognition of emotionally neutral faces. Psychiatry research. 2004;128:123-33.

11. Vikhanova A, Mareschal I, Tibber M. Emotion recognition bias depends on stimulus morphing strategy. Attention, Perception, & Psychophysics. 2022;84(6):2051-9.

12. Dawel A, Miller EJ, Horsburgh A, Ford P. A systematic survey of face stimuli used in psychological research 2000–2020. Behavior Research Methods. 2022;54(4):1889-901.

13. Fernando F-S. Facial Emotion Processing in the Laboratory (and elsewhere): Tradeoffs between Stimulus Control and Ecological Validity. AIMS Neuroscience. 2015;2(4):236-9.

14. Joormann J, Gotlib IH. Is this happiness I see? Biases in the identification of emotional facial expressions in depression and social phobia. J Abnorm Psychol. 2006;115(4):705-14.

15. Paiva-Silva AId, Pontes MK, Aguiar JSR, de Souza WC. How do we evaluate facial emotion recognition? Psychology & neuroscience. 2016;9(2):153.

16. Schurgin MW, Nelson J, Iida S, Ohira H, Chiao JY, Franconeri SL. Eye movements during emotion recognition in faces. Journal of Vision. 2014;14(13):14-.

17. Elfenbein HA, Ambady N. On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological bulletin. 2002;128(2):203.

18. Guarnera M, Magnano P, Pellerone M, Cascio MI, Squatrito V, Buccheri SL. Facial expressions and the ability to recognize emotions from the eyes or mouth: A comparison among old adults, young adults, and children. J Genet Psychol. 2018;179(5):297-310.

19. Testerman M. Face Image Databases: An A-Z directory of databases of face stimulus images for use in behavioral research Princeton University Library2023 [Available from: https://libguides.princeton.edu/facedatabases.

20. Goeleven E, De Raedt R, Leyman L, Verschuere B. The Karolinska Directed Emotional Faces: A validation study. Cognition and Emotion. 2008;22(6):1094-118.

21. Ekman PF, Wallace V. Pictures of Facial Affect. Palo Alto, CA: Consulting Psychologists Press; 1976.

22. Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, et al. The NimStim set of facial expressions: Judgments from untrained research participants. Psychiatry Research. 2009;168(3):242-9.

23. Harris DA, Hayes-Skelton SA, Ciaramitaro VM. What's in a Face? How Face Gender and Current Affect Influence Perceived Emotion. Frontiers in Psychology. 2016;Volume 7- 2016.

24. Wingenbach TSH, Ashwin C, Brosnan M. Sex differences in facial emotion recognition across varying expression intensity levels from videos. PLoS One. 2018;13(1):e0190634.

25. Abrosoft Fantamorph Deluxe 5.4.8 [Available from: https://www.fantamorph.com/index.html.

26. Faceswap [Available from: https://faceswap.dev/.

27. face-parsing.PyTorch [Available from: https://github.com/zllrunning/face-parsing.PyTorch.

28. Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997;10(4):433-6.

29. Sweeny TD, Grabowecky M, Suzuki S, Paller KA. Long-lasting effects of subliminal affective priming from facial expressions. Consciousness and Cognition. 2009;18(4):929-38.

30. Aguado L, Pedraza IS, Gutiérrez AG. A comparison of backward masking of faces in expression and gender identification. Psicologica. 2015;35:171-94.

31. Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. 9th Python in Science Conference; 20102010.

32. Morey RD, Rouder JN. BayesFactor: Computation of Bayes Factors for Common Designs. 2023.

33. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. Behavior research methods. 2009;41:1149-60.

34. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007;9(3):90-5.

35. Waskom ML. seaborn: statistical data visualization. Journal of Open Source Software. 2021;6(60):3021.

36. Forni-Santos L, Osório FL. Influence of gender in the recognition of basic facial expressions: A critical literature review. World J Psychiatry. 2015;5(3):342-51.

37. Cavieres A, Maldonado R, Bland A, Elliott R. Relationship Between Gender and Performance on Emotion Perception Tasks in a Latino Population. Int J Psychol Res (Medellin). 2021;14(1):106-14.