Boschetti, S. et al. (2022). "Eye can't see the difference": Facial Expressions of Pain, Pleasure, and Fear Are Consistently Rated Due to Chance. Human Ethology, 37, 46-72. https://doi.org/10.22330/he/37/046-072

Human Ethology 37(2022): 046-072
Research Article

"Eye can't see the difference": Facial Expressions of Pain, Pleasure, and Fear Are Consistently Rated Due to Chance

Silvia Boschetti[1,2], Hermann Prossinger[3], Tomáš Hladký[1], Kamila Machová[2], Jakub Binter[1,2]

[1]Faculty of Humanities, Charles University, Prague, Czech Republic
[2]Faculty of Science, Charles University, Prague, Czech Republic
[3]Department of Evolutionary Biology, University of Vienna, Vienna, Austria

jakub.binter@fhs.cuni.cz

ABSTRACT

Our research consisted of two studies focusing on the probability of humans being able to perceive the difference between faces expressing pain versus pleasure. As controls, we included: smile, neutral facial expression, and expression of fear. The first study was conducted online and used a large sample (n = 902) of respondents. The second study was conducted in a laboratory setting and involved a stress induction procedure. For both, the task was to categorize whether the facial expression was rated positive, neutral or negative. Stimuli were faces extracted from freely downloadable online videos. Each rating participant (rater) was presented with five facial expressions (stimuli) of five females and of five males. All raters were presented with the stimuli twice so as to evaluate the consistency of the ratings. Beforehand, we tested for stimuli differences using specialized software and found decisive differences. Using a Bayesian statistical approach, we could test for consistencies and due-to-chance probabilities. The results support the prediction that the results are not repeatable but are solely due to chance, decreasing the communication value of the expressions of pain and pleasure. The expression of fear was also rated due to chance, but neither neutral nor smile. Stress induction did have an impact on the perception of pleasure.

Keywords: *perception, emotion, facial expression, visual stimuli, BDSM, pain and pleasure, Dirichlet distribution, Bayesian statistical approach, Cold Pressor Task*

submitted: 03. Sept. 2022; revised: 01. Nov. 2022; accepted: 04. Nov. 2022

INTRODUCTION

Previous studies of stimuli assessment

The ability to estimate, or even deduce, another person's inner feelings and emotions via facial expressions (visual cues) is an essential component of human communication. This ability is particularly important when the communicated facial expression is related to danger, harm, fear, and anger on the one hand and to happiness and surprise on the other (Donato et al., 1999).

Facial expressions have high communicative value for humans; their ability to associate facial expressions with an inner state has been studied in multiple cultures (Ekman 2006; Matsumoto & Kupperbusch, 2001). These studies resulted in *Discrete Category Theory* with seven universal, expressed emotion categories: anger, contempt, disgust, fear, joy, sadness, and surprise. Each of these is considered to be the result of complex grimaces arising from distinct psychophysiological, muscular and neurological activations (Izard, 1994).

A model competing with the aforementioned model, the *Circumplex Model of Emotions*, describes emotions and the associated facial expressions as a two-dimensional phenomenon (rather than the discrete categories of the *Discrete Category Theory*); it is characterized by two perpendicular dimensions, namely valence and arousal (Russell, 1980; Posner at al., 2005). A positive value along the valence axis is often considered the result of the motivation to approach some favorable situation while, conversely, a negative valence is to avoid some unfavorable one. The intensity of an expression is along the arousal axis. The coordinates along these two dimensions are inferred from the observed facial expression; this 2D vector is the emotional assessment attribution.

These two models differ in how to characterize and distinguish each emotion; they make their predictions either by using categories or by quantifying facial expressions. These competing approaches promise straightforward differentiation possibilities. The characterization outcomes need not be equal; we pursue this issue in this manuscript.

Recent studies showed that it is very difficult to correctly identify facial expression of intense emotional states. Aviezer et al. found that the facial expression of intensive states of opposite valence, such as when a tennis player reacts to winning or losing, were not correctly identified by the raters participating in the study (Aviezer et al., 2012). Such counterintuitive observations have been supported by further recent studies (Hughes & Nicholson, 2008; Wenzler et al., 2016). It has, therefore, become evident that humans are not very good in assessing a facial expression displaying an intense emotion in the absence of some contextualization. Contextualization may include auditory (De Gelder & Vroomen, 2000), body posture (Martinez et al., 2016) or some other contextual cue (Zhou & Chen, 2009; Wieser & Brosch, 2012; Kayyal et al., 2015; Alviezer et al., 2017).

There are further influences on perceiving facial expression, such as variabilities of the raters. For example, the biological sex of the rater has been shown to systematically affect the rating of facial expressions (Rotter & Rotter, 1988; Thayer & Johnsen, 2000; Hall & Matsumoto, 2004; Hampson et al., 2006). An evolutionary argument maintains that the facial expression of anger is more easily assessed by males, as they were expected to provide preemptive protection — via anger recognition in an adversary (Rotter & Rotter, 1988). The superior ability by females in other facial expressions has been considered to be adaptive and emerged during our (human) evolutionary history. Because women are the primary caregivers, so the argument, they need to successfully assess the facial expressions of care-receivers. The biological sex of the raters of pain and pleasure expressions has been previously investigated (Hughes & Nicholson, 2008). Their results showed that pain and pleasure expressions are modulated by the biological sex and facial expression of the sender. Female raters showed a better performance in recognizing female expressions of pain. However, in the

Hughes & Nicholson study, the only facial expressions considered were pain and pleasure and no other known universal facial expression (such as fear or joy).

We note that, in previous studies, the facial expressions of differing emotions were portrayed by different male/female faces, therefore preventing investigators to explore the role of the varying expressivities of the persons expressing the emotions via their faces.

Furthermore, the inner state of the individual rating the stimulus can affect the rating. This is true for both arousal and valence (Pell, 2005; Storbeck & Clore, 2008). In several seminal studies, stress was induced by exposure to heights and situations wherein participants expected an electrical shock. A modern — more ethical, yet equally reliable — alternative is to increase physiological arousal by the Cold Pressor Task (Bullinger et al., 1984). It consists of immersing a subject's extremity into ice water for a specified period of time (Mitchell et al., 2004). The sympathetic nervous system is functionally related to the psychological concept of arousal (Dawson et al., 2000) and is responsible for mobilizing the organism's resources to meet internal physiological demands as well as those of the external environment (Salvia et al., 2012).

Studies presented in this paper

Our studies are designed to overcome some of the methodological limitations of previous studies dealing with facial expression of intense affective states.

Aim of Study I: to clarify the consistency of the assessment of such expressions, focusing on the (biological) sex of the rater as well as the biological sexes of the rated (expressers). The stimuli used were images of facial expressions with high (pain, pleasure and fear) and low (neutral and smile) intensity. Previous studies have shown that it is quite easy to recognize smile, fear and neutral and more difficult to recognize pain and pleasure. Each emotion is expressed by each male/female face (they are the two sets of 25 stimuli), so we can control for the sex and the expressivity of the rated individual.

We generate our own set of stimuli by using picture frames from videos that depict consensual acts of extreme sexual activities. We adopted a categorical methodology in which there are three ratings: positive, negative, or neutral. We predicted that, if the individual has been exposed to a negative stimulus (pain, for instance) — the grimace (with expected high intensity), will be rated as negative. In a manifestly opposite stimulus (pleasure, for instance) the rating should be positive.

Aim of Study II: to conduct a follow-up study which evaluated the impact of the inner state of the rating individual. The procedure was similar to Study I with the addition of inducing a stress (Cold Pressure Task) in order to manipulate the inner state of the raters.

METHODS

Sample

Expressers' Faces: In order to be consistent with the published terminology, we use the terms "expressers" and "faces" to describe the individuals who were shown in the 50 video frames as stimuli. We specify the biological sex of the expresser with the terms male and female. The biological sex of the expressers is evident from the video frames.

Raters: In order to be consistent with the published terminology, we use terms "expression raters" and "respondents" to describe the individuals who were presented with the stimuli and who provided their ratings. We specify the biological sex of the expresser with the terms male and female. The (biological) sex we list is the respondent's self-reported one. We deleted all ratings ($n = 4$) of respondents who did not report their biological sex.

In Study I: A total of 902 individuals (aged 18–50; M_{age} = 32 years, SD = 8.9 years) completed the questionnaires; 526 women (M_{age} = 30.9 years, SD = 8.3 years) and 376 men (M_{age} = 33.6 years, SD = 9.5 years). In Study II a total of 28 individuals (aged 19–30; M_{age} = 22.3 years, SD = 2.3 years) took part in the experiment; 13 women (M_{age} = 22.7 years, SD = 2.8 years) and 15 men (M_{age} = 21.9 years, SD = 1.8 years).

Criteria for inclusion were: (a) age of respondents between 18 and 50 years, and (b) at least a minimal experience with adult media, since the facial expressions used in this study were extracted from such materials.

The Two Studies

Study I: The data were collected in the Czech Republic in 2021 via the agency Czech National Panel (narodnipanel.cz) and a science-oriented online portal pokusnikralici.cz using the online platform for data collection Qualtrics®. Participants submitted responses either via computer or mobile devices (smartphones or tablets).

Study II: The data were collected in a laboratory in Prague, Czech Republic. 15 participants were presented with the same stimuli as in Study I; the lower right legs of target group members (n = 15) were immersed in cold water (2–4 °C) for 1½ minutes, which subsequently increased their stress level (Cold Pressor Task; CPT (Bullinger et al., 1984; Brown et al., 2017)). The control group's 13 participants' lower right legs were immersed in water at room temperature.

Stimulus Creation

A method for obtaining the stimuli, as well as their use were presented in a previously published article (Prossinger, 2021b).

From the numerous videos viewed, ten videos (five with female faces and five with male faces) were chosen. Based on the plot in each video, five frames were selected (one of neutrality, one of fear, one of pleasure, one of pain, and one with smile). Three of the authors (S.B. J.B. & T.H.) are researchers in field of human sexuality with more than 10 years of experience, specifically focusing on extreme sexual behavior and on consumption of erotic materials. All three authors (one female and two male) provided their opinion on all of the chosen videos and stimuli choices. All agreed on stimuli choice and what expression is to be expected, based on the contextual information. The agreement on stimuli choice was debated among all three researchers in dedicated meetings.

We point out that it is a common misconception that the individuals taking part in such exchanges derive sexual pleasures from pain and the two happen simultaneously. Although it is not impossible, we have found no mention of this in the published scientific literature. Rather, it should be noted that sensitivity is increased by the feeling of pain by various parts of the body (in our case mainly slapping the buttocks and thighs) and only thereafter is climax achieved. There is no doubt, due to the camera perspective, about the occurrence of the climax in male expressers. In the female expressers no such explicit method of judgement can be used, but all signals of the occurrence of climax were identified by the researchers (involving breathing, contraction of pelvic and anal sphincter muscles, facial blushing, vocalization etc.; Dubray et al., 2017), supported by self-report at the end of the video in some cases.

In each video, male/female faces expressed fear, pain and pleasure during the session, while smile and neutral facial expressions were filmed during an interview prior to the pain and pleasure experiences. All stimuli (images) presented to the raters were scaled to 600 × 600 pixels; we used triangulation between tip of the nose and pupils to ensure that the proportions of the face on the screen were comparable among the frames. No background was visible within the frames presented.

Procedures

In Study I, the set of stimuli was presented twice (Task 1 and Task 2), each time with a different randomization sequence: each stimulus appeared for 1.5 seconds at random intervals ranging from 1 to 3 seconds (so as to avoid constant/rhythmic preparedness for the stimulus presentation). Thus, for each rater, a total of 50 ratings were collected for each presentation set, a total of 100 for both presentations.

In Study II, each rater was presented with the set of stimuli (five male and five female facial expressions) only once. The reason is that the Cold Pressor Task (CPT) has limited impact on the cortisol release and this allowed us to finish the procedure within 20 minutes after the CPT ended.

Ratings

Previous literature (Robertson et al., 2010) has suggested that it is a rather challenging task to correctly identify the facial expression (e.g., to categorize the expression of fear as fear). We therefore asked our participants to rate the observed expression as either positive, neutral, or negative. We thereby avoid the problem of correct labeling and avoided any intricacies associated with a verbal categorization system. The rating was provided by using keyboard keys or a touchpad with dedicated areas with icons.

Statistical Analyses

Due to the inherent advantage of Bayesian statistics when dealing with our research questions, we implemented this approach. General descriptions follow, while more detailed descriptions, augmented by a graphical display, are provided in the Appendix.
(a)Confusion matrices: Both female and male ratings are Dirichlet-distributed (in our case: 3-parametric). The (Bayesian) method of determining whether two groups are significantly different (or not) is to calculate the confusion matrix; it is the obligatory method to use when sample sizes are small. One sample (F) has a distribution $dist_F$ and another sample (G) has a distribution $dist_G$. When there is an overlap of the *pdfs* (probability density functions) of these two distributions, a fraction of F is TRUE (and a fraction is FALSE); likewise, for G. The confusion matrix has four entries:

$$\begin{pmatrix} TRUE_F & FALSE_F \\ FALSE_G & TRUE_G \end{pmatrix}$$

If the off-diagonal elements ({$FALSE_F$, $FALSE_G$}) are small, there exists a significant difference between the distributions of F and of G (the significance level being chosen by the researcher). Observe that the sum of each row in the confusion matrix is 1 = 100%. The fractions in the confusion matrix can also be calculated using Monte Carlo methods.

(b) Possibility of effects being due to chance: In Bayesian statistics, the probability s is a random variable ($0 \leq s \leq 1$). The crucial separator for determining chance is $s = \frac{1}{2}$. The probability is either the integral of the likelihood function $\mathcal{L}(s)$ over the interval $0 \leq s \leq \frac{1}{2}$ or the integral over the interval $\frac{1}{2} \leq s \leq 1$, depending on which side of $s = \frac{1}{2}$ the mode is. In either case, the integral determines whether an observation is due to chance. (A graphical description is shown in the Appendix.) Since there are positive, neutral, and negative responses, we generate a binary case (the correct responses versus the incorrect responses); then the likelihood function is the probability density function of a Beta distribution (see Appendix). For example, for smile, the correct response is the positive rating while the neutral rating and the negative rating together are incorrect responses.

(c) A further way of testing for consistency is by estimating the likelihoods of randomly generated realizations of the two distributions and then using Wilks lambda and its χ^2-distribution in the Laplace limit of large sample sizes.

RESULTS

Correctness Probabilities and Wilks Lambda for Significant Differences

Of the five tested facial expressions, only two were correctly rated with high probability (Table 1) — the stimulus smile as a positive rating and the stimulus neutral as a neutral rating.

Table 1: The three components of the 2 × 5 = 10 modes of the 10 Dirichlet distributions of the ratings of male and female faces for the five expressions by **(a)** female raters and **(b)** male raters. (Note that each mode is the vector $\{mode_A, mode_B, mode_C\}$.) 'Significance' refers to the probability that the male and female face distributions are drawn from the same statistical population. Thus, a significance less than 0.05 means that the two distributions are different at the 5% significance level. **(a)** We observe that female raters rate male faces significantly differently from female faces for smile and for neutral. We also note that smile and neutral are correctly assessed, because, for smile, the mode is de facto on the *A*-axis (positive rating) and very close to 1.00, while for the neutral expression, the mode is far from both *A*- and *C*-axes (therefore mode$_B$ is close to 1.00). In all other cases, the ratings by the females are not consistent with the implied descriptions of the axes. Further discussions are in the text. **(b)** We observe that male raters rate male faces significantly differently from female faces for neutral only. We also note that neutral is correctly assessed, because, for neutral expression, the mode is far from both *A*- and *C*-axes (therefore mode$_B$ is close to 1.00). In all other cases, the ratings by the females are not consistent with the implied descriptions of the axes. Further discussions are in the text.

The probability of a correct rating (a Dirichlet distribution) ranges between 0 and 1 in all cases. The closer to 1 the result (specifically: the component of the mode) is, the higher the probability of correct identification. The remaining probabilities are distributed between the two remaining possibilities.

(a)

Female Raters (Study I)				
Stimulus	Component	Female Faces	Male Faces	Significance
Smile	Positive	0.973	0.946	0.004*
	Neutral	0.021	0.035	
	Negative	0.006	0.019	
Fear	Positive	0.162	0.060	0.5
	Neutral	0.465	0.416	
	Negative	0.373	0.524	
Pain	Positive	0.683	0.515	0.3
	Neutral	0.000	0.000	
	Negative	0.356	0.525	
	Positive	0.451	0.422	

Pleasure	Neutral	0.061	0.148	0.1
	Negative	0.488	0.430	
Neutral	Positive	0.095	0.033	8×10^{-6}*
	Neutral	0.885	0.707	
	Negative	0.020	0.286	

(b)

Male Raters (Study I)				
Stimulus	Component	Female Faces	Male Faces	Significance
Smile	Positive	0.938	0.930	0.3
	Neutral	0.041	0.047	
	Negative	0.021	0.023	
Fear	Positive	0.159	0.127	0.9
	Neutral	0.484	0.483	
	Negative	0.357	0.390	
Pain	Positive	0.612	0.625	0.4
	Neutral	0.000	0.010	
	Negative	0.408	0.365	
Pleasure	Positive	0.408	0.481	0.1
	Neutral	0.069	0.134	
	Negative	0.523	0.385	
Neutral	Positive	0.094	0.095	0.00006*
	Neutral	0.870	0.765	
	Negative	0.036	0.140	

Smile was rated by both sexes with very high accuracy (Table 1). Male participants rated male faces correctly with 0.930 probability and female faces with 0.938 probability; female raters with 0.946 probability for male faces and 0.973 probability for female faces. In the case of female participants rating the smile stimulus, there was a significant difference in rating probability between male faces and female faces (Wilks lambda test; $P < 0.001$) with a higher probability for the latter. For the neutral stimulus, the probability of correct rating by male raters for male faces was 0.765 and 0.870 for female faces. For female participants, the probability of correct rating was 0.707 for male faces and 0.885 for female faces.

The probability of assigning a correct rating for the neutral stimulus was significantly different for male and female faces (Wilks lambda test; $P < 0.001$) in both sexes, with better ratings for female faces.

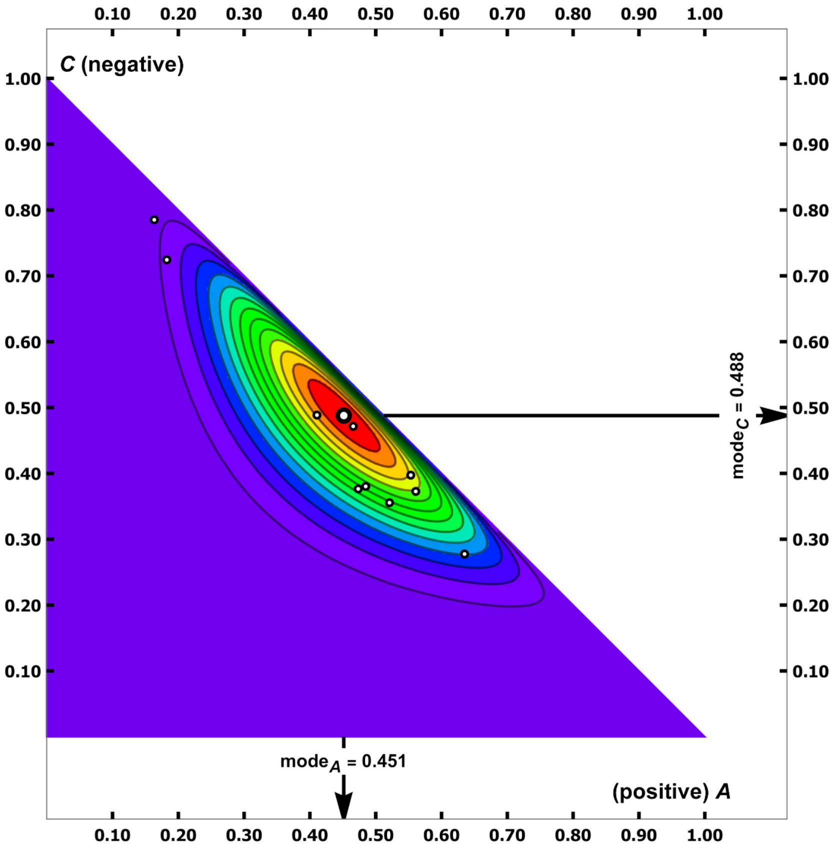
For the three other expressions (pleasure, pain, and fear), the probability of correct rating was very low (Table 1) for raters of both sexes. When rating pleasure, female raters had correctness modes of 0.422 for male faces and 0.451 for female faces; the probabilities of the rating for male faces and female faces were not significantly different; they were equally inaccurate. Male raters (when rating pleasure) had correctness modes of 0.481 for male faces and 0.408 for female faces; these were not significantly different.

When rating pain, female raters had probabilities of rating correctly of 0.525 for male faces and 0.356 for female faces; the ratings were not significantly different. For male raters, the probabilities of rating correctly were 0.365 for male faces and 0.408 for female faces; again not significantly different.

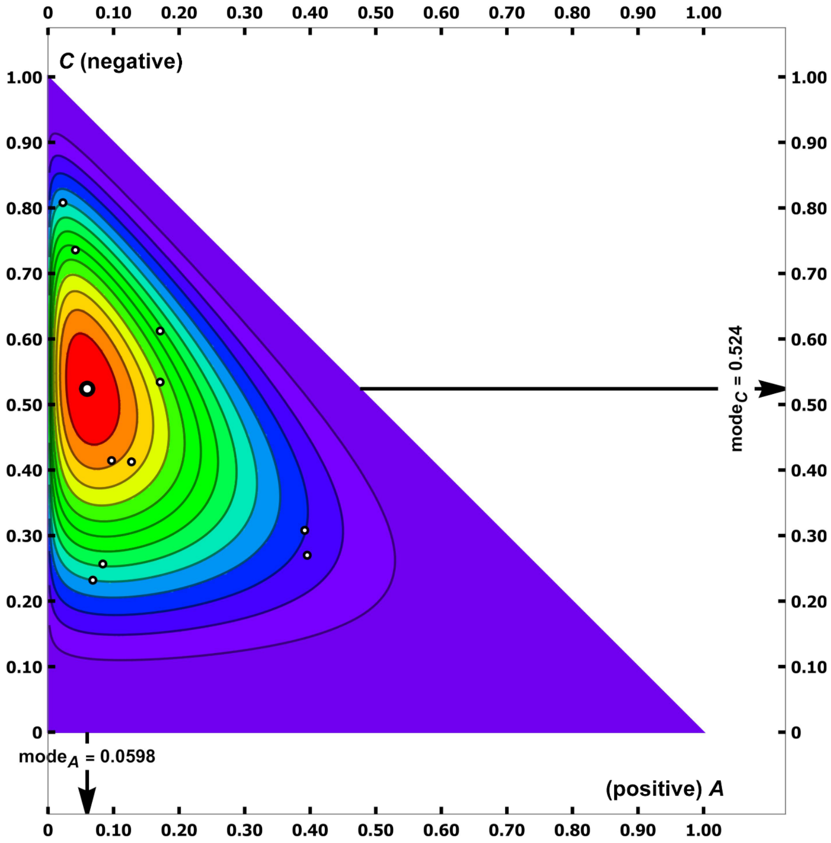
Male raters rated expressions of fear in male faces with 0.390 and in female faces 0.357 probability of correctness. Female raters' correct rating probabilities were 0.524 for male faces and 0.373 for female faces. None of these differences between the ratings of male and female faces were significantly different.

Overall, these results suggest that there is high accuracy in rating of the low arousal expressions, namely neutral and smile. There is small accuracy in the other three facial expressions. The two highly aroused facial expressions (pain and pleasure) have their ratings distributed between the two extreme ratings, namely positive and negative. This is not the case of the fear stimulus where the positive rating probability is very low and the incorrect ratings are towards the neutral rating mode.

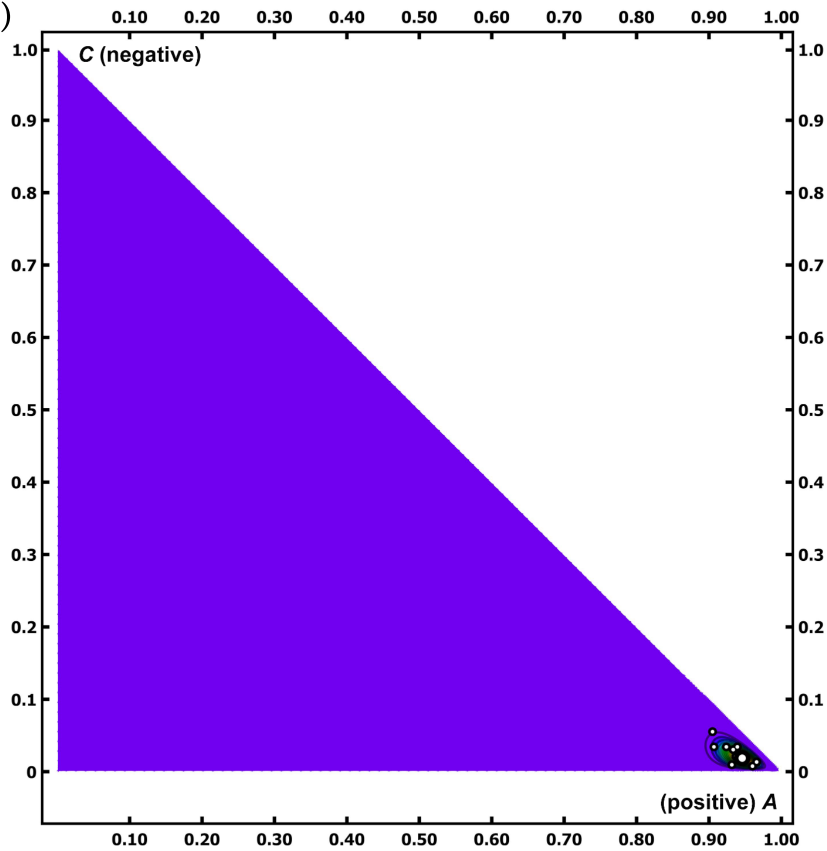
1(a)



1(b)


C (negative)
(positive) A
mode$_C$ = 0.524
mode$_A$ = 0.0598
0
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

1(c)


C (negative)
(positive) A
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

1(d)

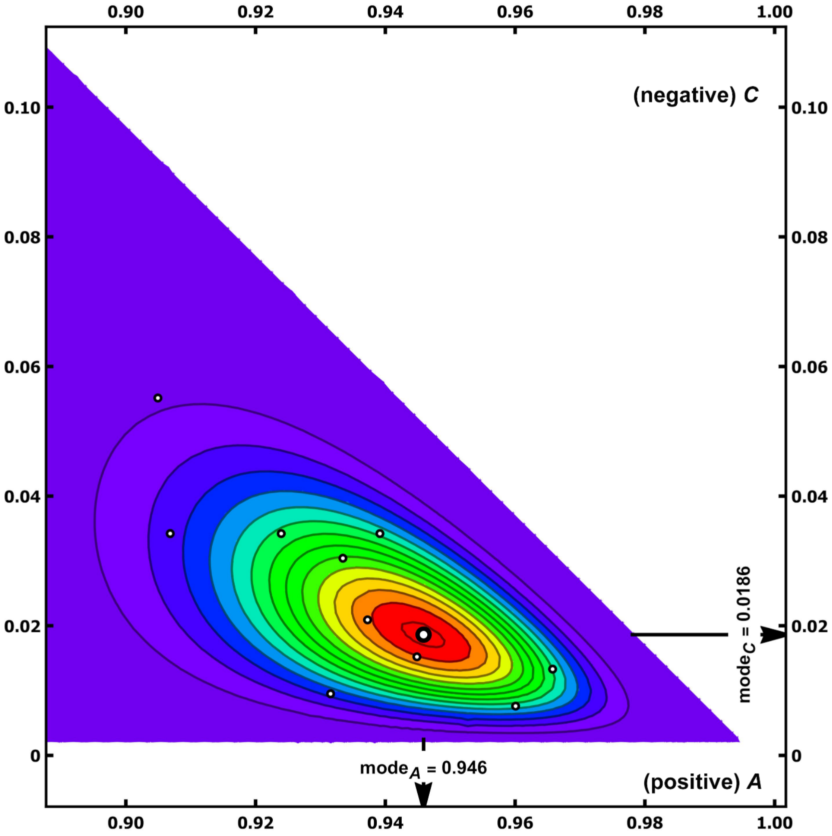


Figure 1: Graphs of three Dirichlet distributions of the ratings by 525 females: **(a)** pleasure twice expressed by females; **(b)** fear twice expressed by males; **(c)** smile twice expressed by males; and **(d)** smile twice expressed by males (detail of (c)). Here, we have chosen s_1 to be the *A*-axis (positive rating) and s_3 to be the *C*-axis (negative rating). The support of the *pdf* of the Dirichlet distribution (shown via likelihood contours) is only defined on a triangle, because $s_1 + s_2 + s_3 = 1$ (therefore, one variable — in our case *B* — cannot be rendered on an axis). The closer the mode is to the hypotenuse, the higher the value of *mode*$_B$. Contours are in steps of $\frac{1}{14}$ the maximum likelihood, color-coded between contours (purple lowest and red highest). The white dots with black borders are the (ten) registration sets (five faces rated twice each) and the large white dot is the mode. **(a)** We observe that the mode component for positive rating is 0.451 (very far from conventionally expected) and the mode component for negative rating is 0.488. Consequently, the mode component for neutral rating is 0.096 (close to conventionally expected). Clearly, pleasure has not been successfully rated by the 525 females; they far too often confused *A* with *C*, showing it is the result of guessing. **(b)** The mode component for positive rating is 0.0598 (as conventionally expected) and the mode component for negative rating is 0524 (very far from conventionally expected). Consequently, the mode component for neutral rating is 0.416 (extremely far from conventionally expected). Clearly, fear has not been successfully rated by the 525 females. **(c)** The mode component for positive rating is 0.973 (as conventionally expected) and the mode component for negative rating is 0.00560 (extremely low, as conventionally expected). Consequently, the mode component for neutral rating is 0.019 (very low, as conventionally expected). Clearly, smile has been successfully rated by the 525 females. **(d)** Detail of (c) showing the numerical values of the smile modes.

An ongoing discussion is about whether women are better at facial expression recognition. This is what we then tested by using a Beta distribution (a Dirichlet distribution with 2 concentration parameters). We have found only two facial expressions where the result was significant. Again, these were the two low arousal expressions (neutral and smile); but only in the case of smile did female raters show a superior ability to categorize the expression (Table 2 and Fig. 1). We note that this difference is negligible since ratings by both sexes were highly accurate (0.935 vs 0.965).

Table 2: The modes of the (Beta) distributions of male and female ratings of the stimuli, separated by sex of the faces. The columns labeled pdtc show the probabilities due to chance. Low probabilities (in the table: $P < 0.001$ — a significance level very far below the conventional $P = 0.05$) show that only the modes of the ratings of stimuli neutral and smile are not due to chance, and these are highly significant. All other modes are due to chance and therefore uninterpretable. The column $\text{sig}_{(\text{male vs female})}$ shows whether the probability that the two (Beta) distributions of male and female facial expressions are significantly different.

Stimulus	Raters	$\text{Mode}_{\text{female}}$	pdtc	$\text{Mode}_{\text{male}}$	pdtc	$\text{sig}_{(\text{male vs female})}$
Smile	male	0.924	<0.001	0.930	<0.001	ns
	female	0.935	<0.001	0.965	<0.001	< 0.002
Fear	male	0.369	>0.999	0.358	>0.999	ns
	female	0.458	>0.999	0.756	>0.999	ns
Pain	male	0.361	>0.999	0.411	>0.999	ns
	female	0.395	>0.999	0.383	>0.999	ns
Pleasure	male	0.463	>0.999	0.409	>0.999	ns
	female	0.412	>0.999	0.445	>0.999	ns
Neutral	male	0.683	<0.001	0.839	<0.001	< 0.001
	female	0.665	<0.001	0.834	<0.001	< 0.002

Heat Maps of individual expresser rating frequencies

Fig. 2 and Table 3 show fractions of participants of both sexes who correctly rated the facial expressions by the individual expressers. The interesting outcome is that there are expressers who are overall better (i.e. inducing a correct rating with a higher probability) than other expressers; furthermore, being better (in the above sense) is not uniformly distributed over all facial expressions.

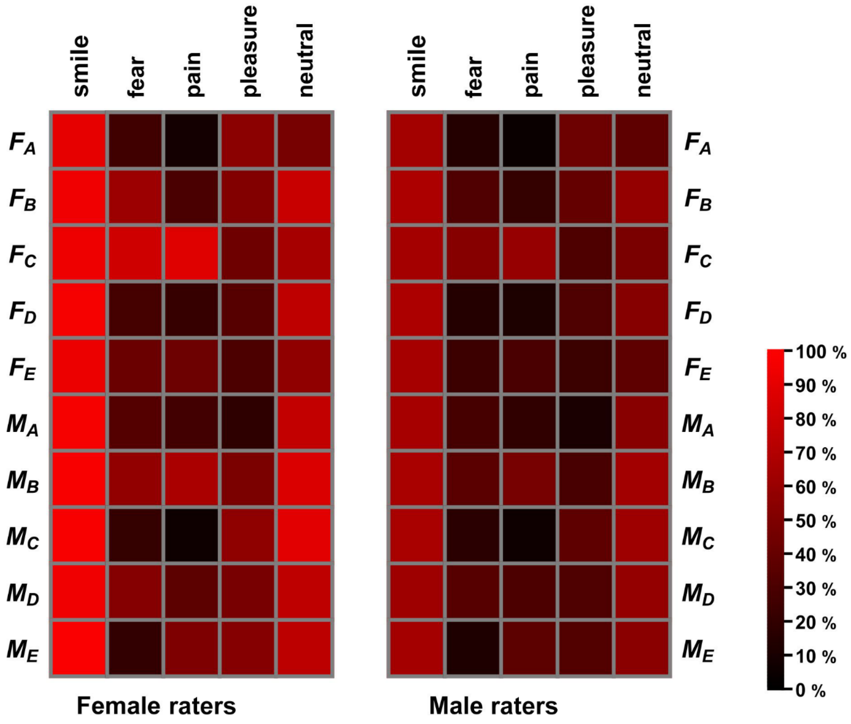


Figure 2: Two heat maps showing the correctness probabilities of ratings by males and females of the male faces and the female faces, face by face. The male faces and female faces are labeled $\mathbf{F_{index}}$ and $\mathbf{M_{index}}$. We observe that female raters correctly rate the expression smile in all female faces and all male faces with very high probability; males somewhat less than females. Remarkably, $\mathbf{F_C}$ and $\mathbf{M_B}$ were rated by the females with a higher correctness probability for fear and pain than were all other female faces and male faces. This phenomenon is comparable, with a lower correctness probability, for the male raters. The females rated all male faces and all female faces with a high probability (70–95%) of correctness for the expression neutral.

Consistencies between Task 1 versus Task 2

Since we presented all stimuli as two consecutively presented tasks, each in a different randomized order, we have the possibility to test the consistency of the ratings. To do so, we have used a Bayesian probability test; the ratings (correct versus incorrect) are Beta distributed. We found one significant result (Table 4), namely only when female raters rated male faces.

Table 3: The modes of the ratings of the Beta Distributions of male and female raters of the five female faces (prefix ‘f’) and the five male faces (prefix ‘m’) expressing the labeled facial expressions during Task 1 and Task 2. The modes are for the probability *s* being correct; if the postulated rating is to be ‘negative’, then *s* is the probability of the raters rating the facial expression as negative.

	Male raters		**Female raters**		**Male raters**		**Female raters**	
	Task 1	**Task 2**	**Task 1**	**Task 2**	**Task 1**	**Task 2**	**Task 1**	**Task 2**
Expresser	*Smile*				*Fear*			
fA	0.912	0.910	0.905	0.939	0.725	0.594	0.754	0.772
fB	0.955	0.910	0.945	0.932	0.655	0.511	0.782	0.757
fC	0.918	0.912	0.933	0.907	0.928	0.914	0.973	0.946
fD	0.957	0.939	0.966	0.960	0.291	0.322	0.406	0.440
fE	0.920	0.904	0.924	0.937	0.733	0.734	0.810	0.764
mA	0.930	0.931	0.968	0.945	0.670	0.639	0.597	0.610
mB	0.941	0.936	0.971	0.983	0.926	0.898	0.938	0.921
mC	0.944	0.960	0.971	0.977	0.500	0.390	0.456	0.397
mD	0.880	0.904	0.945	0.949	0.788	0.774	0.791	0.786
mE	0.915	0.960	0.977	0.968	0.245	0.327	0.269	0.281
Expresser	*Pain*				*Pleasure*			
fA	0.069	0.098	0.090	0.109	0.705	0.630	0.659	0.589
fB	0.458	0.412	0.459	0.548	0.709	0.758	0.680	0.622
fC	0.924	0.917	0.954	0.940	0.522	0.427	0.520	0.372
fD	0.238	0.278	0.330	0.340	0.502	0.492	0.395	0.411
fE	0.455	0.478	0.485	0.527	0.373	0.356	0.355	0.349
mA	0.297	0.187	0.284	0.211	0.166	0.163	0.201	0.172
mB	0.734	0.756	0.739	0.755	0.471	0.728	0.560	0.696
mC	0.087	0.117	0.059	0.079	0.564	0.551	0.601	0.582
mD	0.524	0.583	0.440	0.509	0.526	0.432	0.557	0.457
mE	0.571	0.664	0.543	0.605	0.523	0.432	0.594	0.497
Expresser	*Neutral*							
fA	0.524	0.612	0.470	0.534				
fB	0.816	0.824	0.787	0.814				
fC	0.681	0.758	0.654	0.755				
fD	0.752	0.773	0.743	0.764				
fE	0.527	0.566	0.570	0.557				
mA	0.765	0.777	0.762	0.812				
mB	0.901	0.899	0.859	0.922				
mC	0.883	0.883	0.886	0.930				
mD	0.824	0.840	0.743	0.840				
mE	0.769	0.846	0.736	0.848				

Ratings due to chance

One of the benefits of the Bayesian analytical approach is the possibility to test whether the result obtained is consistent ('real' in common parlance) or if it is obtained due to chance. The probability of the result being due to chance (columns pdtc in Table 2) ranges between 0 and 1; the closer to 1, the more probable that the result is due to chance. The closer the mode (columns Mode_{sex} in Table 2) is to $\frac{1}{2}$, the higher the probability (not: likelihood) of the result being due to chance. For all our results, we obtained extreme ends of the possible outcomes only. Specifically: the two low-arousal expressions (neutral and smile) are not due to chance with probability pdtc < 0.001 (Table 2). In other words, the results are not due to chance at all. A completely opposite result was obtained for the case of the high-arousal faces (fear, pain, and pleasure). The probability is > 0.999, therefore almost certainly due to chance.

Table 4: The consistency of ratings between Task 1 and Task 2. A high significance means that the ratings in Task 1 and Task 2 are consistent. Only one rating (by female raters of male facial expressions — highlighted in light gray) was significantly different between Task 1 and Task 2.

Stimulus	Raters	$\text{Face}_{\text{female}}$	$\text{Face}_{\text{male}}$
Smile	male	> 99%	> 99%
	female	44%	> 99%
Fear	male	> 99%	> 99%
	female	> 99%	> 99%
Pain	male	> 99%	> 99%
	female	> 99%	> 99%
Pleasure	male	> 99%	> 99%
	female	> 99%	17%
Neutral	male	> 99%	> 99%
	female	> 99%	< 1%

Note: The significance of the result is reported in the last two columns.

These results provide further evidence for the above mentioned results, specifically the results of rating the high-arousal stimuli. Not only are the ratings spread between (typically two) options (positive and negative in case of pain and pleasure and negative and neutral in case of fear) but these ratings are also due to chance.

Stress-induced rating differences

In Study II, we analyzed the differences in the distributions of ratings in the two groups of participants (control versus stressed). The confusion matrices (Table 5) display the probabilities of differences of the ratings by the two groups of participants (separately for male faces and female faces). At a 10% significance level (Caelen, 2017), only two off-diagonal elements are significantly different: male faces expressing smile and expressing pleasure were more accurately rated by the stressed group. The shifts (not shown) in correct rating are not due to chance at 5% significance level. (For the *CDF* of the Beta Distributions, the conventional 5% significance level is applicable.)

Table 5: The confusion matrices (entries in %) between the distributions of the ratings by the controlled and the stressed raters, separated by male faces versus female faces. At a significance level of 10% (Caelen, 2017), only two distributions are significantly different (male faces expressing smile and male faces expressing pleasure). A confusion matrix calculation was used instead of the statistical machinery of Wilks lambda, because the sample sizes ($n_{control}$ and $n_{stressed}$) were far from the Laplace condition.

Stimulus	Male Faces	Female Faces
Smile	$\begin{pmatrix} 91.6 & 8.4 \\ 9.7 & 90.3 \end{pmatrix}$	$\begin{pmatrix} 76.3 & 23.7 \\ 18.2 & 81.8 \end{pmatrix}$
Fear	$\begin{pmatrix} 64.3 & 35.7 \\ 30.8 & 69.2 \end{pmatrix}$	$\begin{pmatrix} 75.1 & 24.9 \\ 22.4 & 77.6 \end{pmatrix}$
Pain	$\begin{pmatrix} 57.2 & 42.8 \\ 33.6 & 66.4 \end{pmatrix}$	$\begin{pmatrix} 51.5 & 48.5 \\ 35.1 & 64.9 \end{pmatrix}$
Pleasure	$\begin{pmatrix} 91.5 & 8.5 \\ 7.4 & 92.6 \end{pmatrix}$	$\begin{pmatrix} 84.8 & 15.2 \\ 13.1 & 86.9 \end{pmatrix}$
Neutral	$\begin{pmatrix} 71.2 & 28.8 \\ 27.1 & 72.9 \end{pmatrix}$	$\begin{pmatrix} 73.7 & 26.3 \\ 23.9 & 76.1 \end{pmatrix}$

DISCUSSION

Stimuli Selection

There are multiple ways to produce stimuli for testing. Most often, trained actors or actresses are asked to produce facial expressions that are later rated by professionals or naïve respondents in a pre-test. Whenever the within-rater agreement is sufficiently high, the stimulus was used for testing (an example of this procedure has been published in Kätsyri & Sams, 2008). In our case, this approach is not possible for two reasons: (a) because our hypothesis postulates that the two expressions that are of highest interest to us (pain and pleasure) are putatively indistinguishable, asking pre-test raters to distinguish these would not be sensible, and (b) using stimuli labeled during pre-test as pleasure or pain would inherently lead to testing whether participants agree on representations of pain and pleasure (that is to say, whether there is a common mental representation as discussed by Chen et al., 2018). The ethological validity of such a result would be extremely limited and has been criticized in a recent publication (Van Der Zant & Nelson, 2021). Instead, we followed the methodology of one of the pioneering articles on the topic (Aviezer et al., 2012). The authors searched the internet and chose the stimuli (facial expressions of tennis players) based on the context of winning or losing. The context was not known to their raters but the authors were certain about the outcome of the match and therefore which valence (positive or negative) the stimulus represents.

Following this methodology, we searched for videos online, only including webpages that allowed a free download option. As an extension to the previously mentioned article (Aviezer et al., 2012), we went one step further in our stimuli choice and insisted on finding (and using) individuals of both biological sexes expressing all the five desired expressions.

Ethics of using isolated video frames as stimuli

Although the affliction of pain in the context of extreme, yet consensual, sexual activities is typically due to the actions of some other person, the individual experiencing it knows that he/she consented beforehand and has the ability to demand termination at any time while the scenario evolves; he/she thus retains a high degree of control. Typically, a 'safe-word' is used to terminate such physical and/or psychological activities, and, subsequently, after termination, nurturing behaviors are then provided to such an individual. Pain itself is not an aim of the behavior but rather the goal of increasing sensitivity and priming for greater pleasure (in a sexual sense). It is rare that injuries (apart from bruises) are inflicted on the pain-receiving individual. As stated on the production webpage, all participants in the video clips were informed (not by us, but by the directors) about the to-be-filmed scene contents; they agreed to participation and were interviewed by members of the production team after the scene was completed.

We consider the use of such stimuli as beneficial for science (granted: these stimuli are perceived as controversial by some). They offer the possibility of novel understandings about the problems of the perceptions of facial expressions in several of the evolutionarily most relevant contexts. The acquired knowledge (some of which we have obtained and are presenting in this paper) can, and will certainly be, used in the fields of education, sexuality-related prevention, law enforcement, and therapy. We therefore maintain that the benefits by far outweigh the objections to using such stimuli.

Novelties

Our studies confirmed the results of previous research about the facial expressions of affective states with high arousal — in the absence of further contextual clues. Specifically, the human ability to distinguish between positive and negative valence in cases of facial expressions of extremely high arousal is very weak (Aviezer et al., 2012). Our six design innovations (novelties) allowed us to provide further insights into this topic.

First, because every rater rated each stimulus twice, we could test for consistency. We discovered an increase in accuracy with the second presentation of low arousal stimuli; the increase is ascribable to a recall effect, a learning effect, and a familiarity effect. In the cases of high arousal stimuli, on the other hand, the two ratings appear to be consistent (no difference in accuracy); there cannot be any significant differences, however, because these ratings are due to guessing.

A second novelty is to find the probability of a result being due to chance. To do so, we use Bayesian statistics, which is particularly useful for this challenge. For low arousal stimuli, we find that the result is not due to chance; this infers there must be some mechanism and repeatability is to be expected. In contrast, the three high arousal stimuli (fear, pain and pleasure) are due to chance with an extremely high probability. We find that this is valid for both sexes; we must therefore conclude that discussing any sex-differences is meaningless.

The third novelty is avoiding potential (statistical) noise effects. Every expresser displayed all five facial expressions: fear, smile, pain, pleasure, and neutral.

The fourth novelty is one of design: it deals with every expresser presenting the same five facial stimuli to all male and female raters. We thereby improved (we claim) the reliability of statistical interpretability. Of the biases in data collection (confirmation bias and selection bias), we avoided the latter in this way.

The fifth novelty is the testing of expressions of pain and pleasure not only by women, but also by men.

The sixth novelty deals with a possibility of comparing the ratings of the facial expression of pleasure by males with those by females. The relation between the putative inner feelings of a male when he expresses pleasure facially is much less questionable than for a female. This novelty results in an important departure from the methodology adopted in many other studies of ratings of facial stimuli.

Responses

There have been numerous publications dealing with: (a) differences in facial expression production and degree of expressiveness in various cultures, as well as (b) comparisons of expressiveness of neurotypical versus neurodivergent individuals. Recent publications (Barrett et al., 2019) have concentrated on the individual differences in facial expression production by neurotypical individuals within one culture. We use heat maps (Fig. 2) to display, for the first time, the evidence that raters rate the individual expressers with varying probabilities of success. In other words, we register that some expressers are rated with a higher accuracy for more than one expression whereas others for only one expression.

This phenomenon of some expresser being more accurately rated than others warrants future research, as other circumstances may influence the expressivity of an individual. In naturally occurring (uncontrolled) situations, the strength of the stimulus needed to trigger an affective response varies among individuals. The subsequent research issue is to what degree this affective response triggers a corresponding facial expression. Therefore, the use of naturalistic stimuli results in highly uneven expressions, which necessitate the application of Bayesian statistics.

We did not expect the ratings of fear to be distributed almost equally between negative and neutral. Previous publications provide an explanation: of all the so-called basic expressions, the fearful expression is the least recognizable one, because it is brief and oftentimes admixed with other ones.

Pain and pleasure ratings are almost equally distributed between the extremes positive and negative, with very few neutral ratings. Our results for pain and pleasure ratings are different from those in previous publications.

Human Ratings versus AI Ratings

Our results show that there are no differences between the ratings of pain and pleasure — when rated by humans. We note, however, that there are objective methods for detecting a quantifiable difference in muscle configurations associated with different facial expressions. The recent increase in computation power coupled with the progress in artificial intelligence (AI) techniques provides appropriate tools to test the pain/pleasure rating differences objectively. So there is a difference, but it remains undetectable by humans. In contrast to the publications relying on FACS (Aviezer et al., 2012), we show, in a recently published study (Prossinger, 2021a), how to use an alternative method, based on AI image analysis, to detect objective differences. This algorithmic approach was used to distinguish fearful from neutral faces with a high success rate (Prossinger et al., 2021a). These findings support the existence of an actual difference between two facial expressions. The differences in the expressions are indeed present (and detectable with AI methods) but human raters were unable to detect them with sufficient accuracy.

An interesting comparison, using the same stimuli as in this paper, is provided in another study (Prossinger et al. 2021b); it evaluated the precision of distinguishing stimuli categories. The algorithms found significantly different categories in the case of female expressers. An extension of this research was recently published (Prossinger et al., 2022) with a larger number of female expressers experiencing pain and pleasure. This publication derives important implications about how clustering relates to human raters' inabilities to reliably distinguish between the expressions of pain and pleasure. The study precisely enumerated the far-from-trivial steps necessary for correct classification, which cannot be expected from human vision uncalibrated towards a single individual. There are four clusters and two isolates. These clusters were detected after noise removal. The discovery of the necessity of noise removal provides further support for the two main arguments about the human inability to correctly rate the differences between pain and pleasure. First, the inter-individual facial expression variations are considerable. Second, the (intra-individual) noise component in each specific perception is high. Consequently, it is possible that humans can fine-tune their perception towards certain individuals, especially socially close ones (partners, other family members, or colleagues, for example) thus putatively mitigating noise interference. The AI algorithms are, as shown above, able to overcome this problem.

It is important to point out the inter-individual variability. Even though healthy individuals are all equipped with facial muscles essential for basic emotion expression and the variability of the muscles involved is minimal (Waller et al., 2008), there are many influences related to the uniqueness of each individual´s expressions and limitations in their identification for other individuals in real world scenarios. Some of these limitations are: the fact that people choose or need to wear spectacles, some have beards, some are adorned with jewelry or expensive makeup. All may obstruct or alter the assessment of the facial expression.

Further complications may arise in individuals who experienced facial nerves-related disorders or other central nervous system damage. Furthermore, expressers' age-related features, their fat distribution, their skin texture, their general degree of facial expressiveness, and the morphology of their facial muscles are known to impact the production of their facial expressions (and consequently the probability of correct identification). Therefore, angles and distances between the facial features and their changes from neutral to expressive states constitute the individual expresser's identity (Kande et al., 2000; Yi et al., 2014). We took advantage of this identity uniqueness by using AI algorithms, because the facial identity is rather consistent in adult individuals and it allows for such human-computer interaction (Cohn et al., 2007). Possibly, individual expression familiarity potentially increased the accuracy of correct expression estimation by other humans if they are exposed to an individual for an adequately long, yet unknown in extent, period of time.

In the two studies involving the human raters that are presented in this paper, familiarity was expressly excluded. An interesting next step would be to test such a proposed explanation. Previous studies within this familiarity framework have been conducted on sadness, anger, and happiness; the results are mixed (Zhang & Parmley, 2015). In children, research on pain vocalizations has been published (Corvin et al., 2022); it claimed that learning is the mechanism for obtaining proficiency with respect to specific expressers. It would be worthwhile to compare how successful individuals are in assessing (rating) their partners and relatives in extremely (non-sexually) arousing moments (such as in sports encounters) with the ratings of strangers' facial expressions.

Influence of Arousal on Ratings

A further factor that influences the ratings is that the perception can be affected by the inner state of the rater. We rarely stay calm when encountering highly arousing situations such as winning or losing, reunion with family members, or sexual interaction — in striking contrast to common rating assessment tasks in a research context. These situations involve emotional coupling and affect mirroring and cause a dynamic attribution process (Hasson & Frith, 2016). Indeed, the state of the rater affects the perception of the expresser.

Dutton and Aron (1974) made their participants rate ambiguous pictures with the Thematic Apperception Test (TAT). Those with induced anxiety rated the situations in the pictures as having increased sexual connotations. If we assume that pain-pleasure is equally ambiguous as the stimuli (pictures) used in the TAT, we would predict a shift towards a more positive rating for both pain and pleasure. Brown et al., (2017) were among the first to design a comparable test by using the Cold Pressor Task to find a possible shift towards the negative rating in the case of a surprised face — which was considered ambiguous by those authors. Three facial expressions had been presented: happy face, surprise face, and angry face. The happy face and the angry face were not affected by the induced stress. Those stressed rated the surprise face as more negative.

Our Study II is a more refined version of Brown et al.'s study, because we used five facial expressions comprising 50 stimuli. Also, our analysis is based on Bayesian statistics, which avoids sample size issues and allows for further insights, notably due-to-chance probabilities.

The outcomes of our Study II, in which we manipulated the arousal of the raters, document no shift in rating at 5% significance level. The observed shift towards more positive rating only happened in the cases smile and pleasure in male expressers at 10% significance. The choice of significance levels in confusion matrices is dynamic; research indicates that the choice 5% is rarely warranted; 10% is to be preferred (Caelen, 2017).

Implications

Facial expressions of pain, pleasure and fear are uninformative. Because the display (of pain and pleasure) is ambiguous, the signal perceived by the rater is uninformative. The misinformation can be exacerbated by arousal change in the rater.

We therefore consider an implication to be: verbal communication is a practical resolution of the above ambiguities during many (but perhaps not all) interactions in real life.

LIMITATIONS AND FUTURE DIRECTIONS

One seeming limitation is the prediction of null results. In a statistical sense, it is considered problematic to test for a null effect (but that is perhaps due to Null Hypotheses Statistical Testing conventions and the associated fallacies). Bayesian statistics is not susceptible to such a problem (because the method does not violate Bayes' Theorem) and specifically includes testing for a null result. Therefore, this approach is promising for future research.

We tested for two types of null result. One null result (often observed): the outcome of a statistical test shows that the observed effect is due to chance. The other type we tested for: that the observed difference of a result that is not due to chance but the detected difference is valid with a very small probability.

In both studies presented here, the samples of both stimuli and raters consisted of members of a Caucasian population, since the diversity of population in the Czech Republic is minimal. The results, although very strong, may not be directly generalizable to other populations.

Female sexual pleasure is difficult to assess; but this difficulty applies to all related research. There are claims that even self-reports would not be sufficient. Devices used for measuring female sexual arousal are insufficiently reliable (Meston et al., 2004; Cooper et al., 2014; Meston et al., 2019), so we cannot rely on their applicability in this investigation. As in other studies that attempt to relate arousal with female pleasure expression, we use the pragmatic approach: for stimulus creation, it is sufficient to adopt the convention of relying on using already existing, freely downloadable videos. Researchers who question this pragmatic approach must then reject the validity of a vast number of studies dealing with facial expression of pleasure, not only those using videos. However, it should be pointed out that applying the AI methods to facial expressions (Prossinger et al., 2022) have the potential of resolving this impasse.

By the same token, we feel the need to address the possibility that the expression seen does not match the inner feeling of the expresser. This is not a design flaw but involves an inherently biological aspect in the field of research using naturalistic stimuli.

Furthermore, the expression of fear as a reliable stimulus may be considered problematic since the expressers were aware of the fact that, ultimately, the situation is safe: no permanent damage is de facto guaranteed. Fear, of all expressions considered basic, has the lowest identification reliability rate, and this is especially true in naturalistic expression scenarios. In other words, the results obtained are less unusual than may appear at first glance.

Lastly, it should be pointed out that the situation of sexual play is not transferable to other types of interaction where such mismatches can be found, e.g., sport, fighting, injury infliction. Therefore, generalization of our findings to such fields should be used with caution.

CONCLUSIONS

Recent studies dealing with facial expressions are shifting from laboratory-produced situations with pre-tested expressions towards the more real-world relevant way of stimuli creation in order to obtain context-dependent facial expressions. Due to the change in study design, the observed outcomes are remarkably different. These different outcomes challenge many of the cornerstones of this research field. Furthermore, our study is currently unique in repeatedly using one expresser's face for all five expressions. In other words, participants were presented with 50 individual stimuli, one at a time, of five different individuals expressing five different grimaces (namely smile, fear, pain, pleasure, and neutral). We find that, for human raters, perception of facial expressions of pain and pleasure are ambiguous. The participants were specifically requested to supply a categorical rating so as to avoid errors related to descriptive ratings. All insights were obtained by using a Bayesian statistical approach, which also allows for testing probabilities due to chance and a reliability measure for a null result.

The results for low-arousal expressions (smile and neutral) confirm that the method and the analytical approach are appropriate for investigating the observations. The low-arousal expressions were rated with high accuracy and with high probability, inferring that these are repeatable results. Our findings regarding high-arousal expressions, on the other hand, confirm that, even though there are objective differences in the expressions of pain and pleasure (which were tested using AI methods), they are indistinguishable by humans, especially when trying to ascertain such differences in strangers.

The rating options for pain and pleasure are actually binary (positive or negative). Even when offering a neutral distractor, we find that the ratings are always due to chance. Furthermore, this result is repeatable, which we tested by two presentations of each stimulus. In other words, guessing is the only reason for a null rating — and we did not find a learning effect. This disqualifies further analyses regarding statistical variabilities among the raters. Similar results were obtained for the fear expression. Thus, the ratings were predominantly distributed between the negative and neutral category; rarely was fear rated as positive. However, all these ratings were also due to chance.

To our knowledge, this is the first time that the ambiguous facial expressions of one expresser were presented to participants who were also in a condition of increased arousal. This procedure shifted ratings to more accurate ones, we found, namely for two positive facial expressions (smile — a low arousal expression — and pleasure — high arousal expression) of male expressers. The other expression ratings were unaffected by arousals induced in the raters. This could suggest that there is, with arousal, a selective shift in the positive expression perception: it is in concordance with the original work on misattribution of arousal.

ETHICS

Even though the materials presented to the participants were not *per se* of a sexual nature (as only facial expressions were presented) we made precautions to limit any negative impact on our participants.

Informed consent

In Study I: An online information text and consent form was supplied; after reading it, a box was to be ticked by each participant (indicating their informed consent) prior to their participation.

In Study II: Two informed consent forms were to be manually/personally signed. The first was presented to a potential rater prior to participation; it included all the information about procedures (including the CPT), safety measures, kinds of data collected, and risks. The second informed consent form consisted of a full disclosure of the aim(s) of the study, the expected impact of the procedures, and the possible implications for the rater signing this second form. It was to be signed after the debriefing procedure (see below). If the second consent form was not signed, the collected data was discarded (and therefore not used in the analysis).

Post-study Support and Debriefing

All parts of the design and debriefing were conducted in co-operation with a trained psychologist who also supervised all data collection.

For Study I we supplied the participants with a list of contacts: (1) to the principal investigator, (2) to a psychological counseling center, and (3) to an organization that deals with sexuality-related issues.

During the debriefing phase for Study II, every rater participated in a debriefing discussion by a trained psychologist directly after the completion of data collection. The rater then received a written detailed description, with a full explanation of the possible negative aspects of the experiment, especially those related to the stress-induction procedure, and was also supplied with a list of contacts: (1) to the principal investigator, (2) to a psychological counseling center, and (3) to an organization that deals with sexuality-related issues.

ACKNOWLEDGMENTS

The study was preregistered on OSF portal: https://osf.io/bhk6m/.

FUNDING

This research was funded by the Czech Science Foundation in the project with number GACR 19-12885Y and titled “Behavioral and Psycho-Physiological Response on Ambivalent Visual and Auditory Stimuli Presentation”.

INSTITUTIONAL REVIEW BOARD STATEMENT

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Faculty of Science, Charles University, Prague, Czech Republic (Protocol Code 2018/08, approval date: 2 April 2018).

DATA AVAILABILITY STATEMENT

The data are available on the OSF portal.

The frames were extracted from commercially available online videos. As the videos are proprietary, we can only make the extracted frames we used available from the corresponding author (upon reasonable requests originating from a serious institutional email address).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science, 338*(6111), 1225–1229. DOI

Aviezer, H., Ensenberg, N., & Hassin, R. R. (2017). The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology, 17,* 47–54. DOI

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68. DOI

Brown, C. C., Raio, C. M., & Neta, M. (2017). Cortisol responses enhance negative valence perception for ambiguous facial expressions. *Scientific Reports, 7*(1), 1–8. DOI

Bullinger, M., Naber, D., Pickar, D., Cohen, R. M., Kalin, N. H., Pert, A., & Bunney Jr, W. E. (1984). Endocrine effects of the cold pressor test: relationships to subjective pain appraisal and coping. *Psychiatry Research, 12*(3), 227–233. DOI

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence, 81*(3), 429–450. DOI

Chen, C., Crivelli, C., Garrod, O. G., Schyns, P. G., Fernández-Dols, J. M., & Jack, R. E. (2018). Distinct facial expressions represent pain and pleasure across cultures. *Proceedings of the National Academy of Sciences, 115*(43), E10013–E10021. DOI

Cohn, J. F. (2007). Foundations of human computing: Facial expression and emotion. In: *Artifical Intelligence for Human Computing* (pp. 1-16). Springer, Berlin, Heidelberg. DOI

Cooper, E. B., Fenigstein, A., & Fauber, R. L. (2014). The faking orgasm scale for women: Psychometric properties. *Archives of Sexual Behavior, 43*(3), 423-435. DOI

Corvin, S., Fauchon, C., Peyron, R., Reby, D., & Mathevon, N. (2022). Adults learn to identify pain in babies' cries. *Current Biology, 32*(15), R824–R825. DOI

Dawson, M., Schell, A., Filion, D. (2000). The electrodermal system. *Handbook of Psychophysiology*. Cambridge University Press. DOI

De Gelder, B. & Vroomen, J. (2000) The perception of emotions by ear and by eye. *Cognition & Emotion 14,* 289–311. DOI

Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J. (1999). Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(10), 974–989. DOI

Dubray, S., Gérard, M., Beaulieu-Prévost, D., & Courtois, F. (2017). Validation of a self-report questionnaire assessing the bodily and physiological sensations of orgasm. *The Journal of Sexual Medicine, 14*(2), 255-263. DOI

Dutton, D. G., & Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology, 30*(4), 510. DOI

Ekman, P. (2006). *Darwin and facial expression: A century of research in review.* NY: Academic Press

Hampson, E., van Anders, S. M., & Mullin, L. I. (2006). A female advantage in the recognition of emotional facial expressions: Test of an evolutionary hypothesis. *Evolution and Human Behavior, 27*(6), 401–416. DOI

Hall, J. A., & Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion, 4*(2), 201. DOI

Hasson, U., & Frith, C. D. (2016). Mirroring and beyond: coupled dynamics as a generalized framework for modelling social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1693), 20150366. DOI

Hughes, S. M. & Nicholson, S. E. (2008). Sex differences in the assessment of pain versus sexual pleasure facial expressions. *Journal of Social, Evolutionary, and Cultural Psychology* 2(4), 289. DOI

Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*(2), 288–299. DOI

Kanade, T., Cohn, JF, & Tian, Y. (2000). Comprehensive database for facial expression analysis. *Proceedings fourth IEEE international conference on automatic face and gesture recognition* (cat. No. PR00580), pp. 46-53. DOI

Kätsyri, J., & Sams, M. (2008). The effect of dynamics on identifying basic emotions from synthetic and natural faces. *International Journal of Human-Computer Studies, 66*(4), 233–242. DOI

Kayyal, M., Widen, S., & Russell, J. A. (2015). Context is more powerful than we think: Contextual cues override facial cues even for valence. *Emotion, 15*(3), 287. DOI

Matsumoto, D., & Kupperbusch, C. (2001). Idiocentric and allocentric differences in emotional expression, experience, and the coherence between expression and experience. *Asian Journal of Social Psychology, 4*(2), 113–131. DOI

Martinez, L., Falvello, V. B., Aviezer, H., & Todorov, A. (2016). Contributions of facial expressions and body language to the rapid perception of dynamic emotions. *Cognition and Emotion, 30*(5), 939–952. DOI

Meston, C. M., Levin, R. J., Sipski, M. L., Hull, E. M., & Heiman, J. R. (2004). Women's orgasm. *Annual Review of Sex Research, 15*(1), 173–257.

Meston, C. M., & Stanton, A. M. (2019). Understanding sexual arousal and subjective-genital arousal desynchrony in women. *Nature Reviews Urology, 16*(2), 107–120. DOI

Mitchell, L. A., MacDonald, R. A., & Brodie, E. E. (2004). Temperature and the cold pressor test. *Journal of Pain, 5*(4), 233–237. DOI

Pell, M. D. (2005). Nonverbal emotion priming: evidence from the 'facial affect decision task'. *Journal of Nonverbal Behavior, 29*(1), 45–73. DOI

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology, 17*(3), 715–734. DOI

Prossinger, H., Binter, J., Hladký, T., & Říha, D. (2021a). Using Neural-Network-Driven Image Recognition Software to Detect Emotional Reactions in the Face of a Player While Playing a Horror Video Game. *International Conference on Human-Computer Interaction* (pp. 258–265). Springer, Cham. DOI

Prossinger, H., Hladky, T., Binter, J., Boschetti, S., & Riha, D. (2021b). Visual Analysis of Emotions Using AI Image-Processing Software: Possible Male/Female Differences between the Emotion Pairs "Neutral"–"Fear" and "Pleasure"–"Pain". *The 14th PErvasive Technologies Related to Assistive Environments Conference* (pp. 342–346). DOI

Prossinger, H., Hladký, T., Boschetti, S., Říha, D., & Binter, J. (2022). Determination of "Neutral"–"Pain","Neutral"–"Pleasure", and "Pleasure"–"Pain" Affective State Distances by Using AI Image Analysis of Facial Expressions. *Technologies, 10*(4), 75. DOI

Roberson, D., Damjanovic, L., & Kikutani, M. (2010). Show and tell: The role of language in categorizing facial expression of emotion. *Emotion Review, 2*(3), 255–260. DOI

Rotter, N. G., & Rotter, G. S. (1988). Sex differences in the encoding and decoding of negative facial emotions. *Journal of Nonverbal Behavior, 12*(2), 139–148. DOI

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161. DOI

Salvia, E., Guillot, A., & Collet, C. (2012). Autonomic nervous system correlates to readiness state and negative outcome during visual discrimination tasks. *International Journal of Psychophysiology , 84*(2), 211–218. DOI

Storbeck, J., & Clore, G. L. (2008). Affective arousal as information: How affective arousal influences judgments, learning, and memory. *Social and Personality Psychology Compass, 2*(5), 1824–1843. DOI

Thayer, J., & Johnsen, B. H. (2000). Sex differences in judgement of facial affect: A multivariate analysis of recognition errors. *Scandinavian Journal of Psychology, 41*(3), 243–246. DOI

Van Der Zant, T., & Nelson, N. L. (2021). Motion increases recognition of naturalistic postures but not facial expressions. *Journal of Nonverbal Behavior, 45*(4), 587–600. DOI

Waller, B. M., Cray Jr, J. J., & Burrows, A. M. (2008). Selection for universal facial emotion. *Emotion, 8*(3), 435. DOI

Wenzler, S., Levine, S., van Dick, R., Oertel-Knöchel, V., & Aviezer, H. (2016). Beyond pleasure and pain: Facial expression ambiguity in adults and children during intense situations. *Emotion, 16*(6), 807. DOI

Wieser, M. J. & Brosch, T. (2012) Faces in context: a review and systematization of contextual influences on affective face processing. *Frontiers in psychology 3,* 471. DOI

Yi, J., Mao, X., Chen, L., Xue, Y., & Compare, A. (2014). Facial expression recognition considering individual differences in facial structure and texture. *IET Computer Vision, 8*(5), 429-440. DOI

Zhang, F., & Parmley, M. (2015). Emotion attention and recognition of facial expressions among close friends and casual acquaintances. *Journal of Social and Personal Relationships, 32*(5), 633–649. DOI

Zhou, W. & Chen, D. (2009). Fear-related chemosignals modulate recognition of fear in ambiguous facial expressions. *Psychological Science, 20,* 177–183. DOI

APPENDIX

Dirichlet Distribution

The ratings by female raters are Dirichlet distributions (in our case with three concentration parameters $\{a_A, a_B, a_C\}$), as are those of the males. We predicted the repeats (Trial I versus Trial II) to be the same, and we tested for that. We therefore have, for female raters rating five female faces displaying fear, ten registration sets with triples $\{n_A, n_B, n_C\}$ in each set, with $n_A + n_B + n_C = 10$. The *pdf* (probability density function) of the Dirichlet distribution $\mathcal{Dir}$, called the likelihood function $\mathcal{L}(s_1, s_2, s_3) = pdf(\mathcal{Dir}(a_A, a_B, a_C), s_1, s_2, s_3)$ with concentration parameters $\{a_A, a_B, a_C\}$ and probabilities s_1, s_2, s_3 of observing the variables var_1, var_2, var_3 is

$$\mathcal{L}(s_1, s_2, s_3) = pdf(\mathcal{Dir}(a_A, a_B, a_C), s_1, s_2, s_3) = \frac{\Gamma(\alpha_A + \alpha_B + \alpha_C)}{\Gamma(\alpha_A)\Gamma(\alpha_B)\Gamma(\alpha_C)} s_1^{\alpha_A - 1} s_2^{\alpha_B - 1} s_3^{\alpha_C - 1}$$

with $s_3 = 1 - s_1 - s_2$ and $0 \leq s_i \leq 1\ \forall i\ i = 1 \dots 3$; $\Gamma(\cdots)$ is the Gamma function.

The two modes for A and C are $mode_A = \frac{(\alpha_A - 1)}{(\alpha_A + \alpha_B + \alpha_C - 3)}$ and $mode_C = \frac{(\alpha_C - 1)}{(\alpha_A + \alpha_B + \alpha_C - 3)}$.

If we are interested in axes A and B, rather than A and C, then the formulae are cycled. Below, we explain why we use which axes and when. Note that the formulae for the modes are straightforward, suggesting we need not use the (somewhat complicated) formula for the probability density function *pdf*. However, there are no closed algebraic formulas for the uncertainty intervals for the *pdf* of the Dirichlet, but there are contours of uncertainty (see, for example, Fig. 1), and these contours are oddly-shaped smooth curves. We need to analyze the contour geometry in order to interpret possible overlap (which enables us to determine whether the rating distributions of male and female raters are significantly different).

Bayesian estimation of guessing

Each face is rated as exhibiting one of the five facial expressions. We do not expect, but do postulate — as a test — that the facial expression smile (for example) will be rated positive, while the facial expression pain will be rated negative. We use a Bayesian approach to determine the maximum likelihood of a correct probability(!). For each face of each facial expression rated by the females (say), let n_1 be the number of ratings that agree with the postulated rating, while n_2 is the number of ratings that disagree with the postulated rating (then $n_1 + n_2 = n$; $n = 526$ for female raters; $n = 376$ for male raters). In Bayesian statistics, in which the probability s is a random variable, the likelihood function, for this situation, $pdf(s) = \mathcal{L}(s)$ of s is a Beta Distribution

$$\mathcal{L}(Be(\alpha, \beta), s) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} s^{\alpha - 1}(1 - s)^{\beta - 1} = \frac{\Gamma(n_1 + n_2 + 2)}{\Gamma(n_1 + 1)\Gamma(n_2 + 1)} s^{n_1}(1 - s)^{n_2}$$

The probability (in Bayesian statistics) of observing a result disagreeing with the postulate is then,

$$\int_0^{1/2} \mathcal{L}(Be(\alpha, \beta), s) ds$$

The most likely probability s_{ML} is the mode. s_{ML} = mode = $\frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$. We note that the postulate is always s, even if the postulated rating is negative (as in the case of pain).

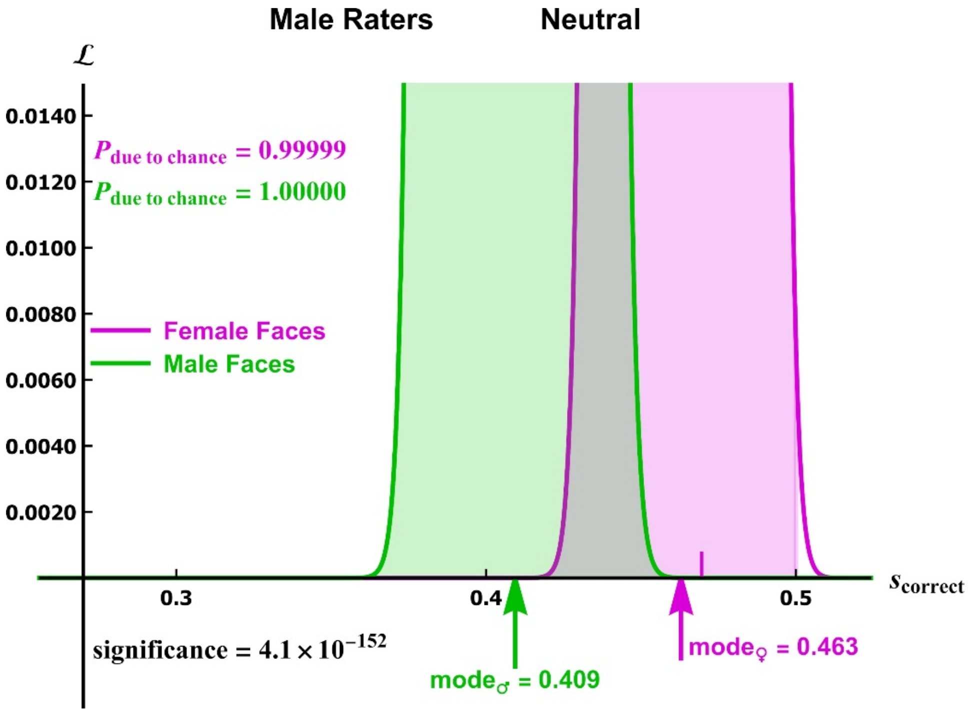


Figure A-1: An example of the relationships between the independence of two samples and the determination of an outcome due to chance. The likelihood functions of the Beta distributions of the two samples are shown, along with the areas under the curve (shaded); these areas show the probability of the observed distribution of the ratings being due to chance (i.e. guessing by the male raters of the presented stimulus). One is $P_{\text{due to chance}}$ = 1.000 … (very high probability that the raters are guessing), the other is $P_{\text{due to chance}}$ = 0.99999 … (again, very high probability that the raters are guessing; in this latter case, the small unshaded area clarifies why the probability is close, but not equal to, 1.0000…). The modes are significantly different, because the probability that the two observed rating sets are drawn from the same statistical population is 4.1×10^{-132}; in other words, it is extremely unlikely that the two distributions are samplings from the same statistical population. This significance has been calculated with Wilks lambda (see below). The peaks of the likelihood functions are not shown; the details near the s_{correct}-axis have been shown.

An example of a result is shown in Fig. A-1. For each rating set (male or female) of all 10 faces, we obtain, for each expression, two modes, one for Task 1 and one for Task 2.

Testing for independence of two distributions: Wilks Lambda

Given: two samples of ratings (of the stimulus pleasure, say), one by females (total counts n_F with n_G correct) and one by males (total counts n_M with n_H correct). The distributions are Beta distributions. If $s = s_{\text{correct}}$ is the probability of a correct rating, then the likelihood function is, for the females,

$$\mathcal{L}_F(s) = \frac{\Gamma(n_F + 2)}{\Gamma(n_H + 1)\Gamma(n_F - n_G + 1)} s^{n_G}(1 - s)^{n_F - n_G}$$

and, for the males,

$$\mathcal{L}_M(s) = \frac{\Gamma(n_M + 2)}{\Gamma(n_H + 1)\Gamma(n_M - n_H + 1)} s^{n_H}(1 - s)^{n_M - n_H}$$

We generate random numbers N_A using the female likelihood function and random numbers N_B using the male likelihood function. We then estimate the ML Beta distribution $dist_\mathrm{A}$ using the N_A random numbers and the ML Beta distribution $dist_\mathrm{B}$ using the N_B random numbers. We also estimate the ML distribution $dist_\mathrm{AB}$ of the combined random numbers $N_\mathrm{AB} = N_\mathrm{A} \cup N_\mathrm{B}$. We calculate the three log-likelihoods

$$\mathrm{logLike}_\mathrm{A} = \ln\mathcal{L}(dist_\mathrm{A}|N_\mathrm{A})$$

$$\mathrm{logLike}_\mathrm{B} = \ln\mathcal{L}(dist_\mathrm{B}|N_\mathrm{B})$$

$$\mathrm{logLike}_\mathrm{AB} = \ln\mathcal{L}(dist_\mathrm{AB}|N_\mathrm{AB})$$

and Wilks Lambda

$$\Lambda_{Wilks} = -2(\mathrm{logLike}_\mathrm{AB} - (\mathrm{logLike}_\mathrm{A} + \mathrm{logLike}_\mathrm{B})),$$

which is, in the Laplace/frequentist paradigm (i.e. *sample size* → ∞), χ^2-distributed with df = $\mathrm{df}_\mathrm{AB} - (\mathrm{df}_\mathrm{A} + \mathrm{df}_\mathrm{B})$ degrees of freedom (df = 2 in our analysis). Thus, if significance = $\mathrm{CDF}(\chi^2(2), \Lambda_{Wilks})$ is very small, we conclude that the probability that the two samples with their respective Beta Distributions are improbably drawn from the same statistical sample.

Fig. A-1 shows an example. The significance is very small, so we conclude that the probability that the two samples (male and female faces rated by male raters) are drawn from the same statistical population is significantly small; therefore, the modes are significantly different. However, both the distributions of rating samples cannot be excluded from being due to chance.