Journal of Experimental Social Psychology 103 (2022) 104394

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

The effect of face masks on the stereotype effect in emotion perception☆

Maximilian A. Primbs *, Mike Rinck, Rob Holland, Wieke Knol [1], Anique Nies [1], Gijsbert Bijlstra

Behavioural Science Institute, Radboud University, Thomas van Aquinostraat 4, 6525GD Nijmegen, the Netherlands

ARTICLE INFO

Keywords:
Stereotypes
Emotion perception
Face masks
Specification curve analysis

ABSTRACT

The accurate and swift decoding of emotional expressions from faces is fundamental for social communication. Yet, emotion perception is prone to error. For example, the ease with which emotions are perceived is affected by stereotypes (Bijlstra, Holland, & Wigboldus, 2010). Moreover, the introduction of face masks mandates in response to the Covid-19 pandemic additionally impedes accurate emotion perception by introducing ambiguity to the emotion perception process. Predictive coding frameworks of visual perception predict that in such situations of increased ambiguity of the sensory input (i.e., faces with masks), people increasingly rely on their prior beliefs (i.e., their stereotypes). Using specification curve analysis, we tested this prediction across two experiments, featuring different social categories (Study 1: Gender; Study 2: Ethnicity) and corresponding emotion stereotypes. We found no evidence that face masks increase reliance on prior stereotypes. In contrast, in Study 1 (but not in Study 2), we found preliminary evidence that face masks decrease reliance on prior stereotypes. We discuss these findings in relation to predictive coding frameworks and dual process models and emphasize the need for up-to-date analytic methods in social cognition research.

1. Introduction

The accurate and swift decoding of emotional expressions from human faces is fundamental for social communication. As such, failures and deficits in facial emotion perception can have negative consequences, such as poor social functioning, decreased quality of social interactions, and inappropriate behavioral responses (e.g., in autism: Baron-Cohen, Richler, Bisarya, Gurunathan, & Wheelwright, 2003). Successful emotion perception depends on a multitude of factors, such as the facial features of the observed (Becker, Kenrick, Neuberg, Blackwell, & Smith, 2007; Marsh, Adams Jr., & Kleck, 2005; Sacco & Hugenberg, 2009), neurological disorders of the observer (e.g., Ashwin, Chapman, Colle, & Baron-Cohen, 2006; Harms, Martin, & Wallace, 2010) or evaluative associations between social categories and emotions (Craig, Koch, & Lipp, 2017; Hugenberg, 2005; Hugenberg & Sczesny, 2006). In an example of the latter, Hugenberg (2005) presented White observers with White and Black faces displaying different emotions. He found that the observers correctly identify happiness comparatively faster on White faces, and anger and sadness comparatively faster on Black faces and concluded that the race of the target face provides an evaluative context in which emotions are interpreted. This suggests that White observers have more positive evaluative associations with White faces compared to Black faces and more negative evaluative associations with Black faces compared to White faces.

Other studies argued that social categories do not merely influence emotion perception via general negative or positive evaluative associations, but also via specific stereotype associations (Bijlstra et al., 2010; Bijlstra, Holland, Dotsch, & Wigboldus, 2019), a claim well in line with prior research differentiating both processes (Amodio & Devine, 2006). In their research, Bijlstra et al. (2010) employed a speeded categorization task comparing the time required to perceive anger or sadness on male and female faces and on White-Dutch and Moroccan-Dutch male faces. Congruent with the gender stereotype that anger is more typical of males than of females and sadness more typical of females than of males (Plant, Hyde, Keltner, & Devine, 2000), and the in the Netherlands widespread stereotype association between Moroccan-Dutch males and danger (Bijlstra, Holland, Dotsch, Hugenberg, & Wigboldus, 2014; Dotsch & Wigboldus, 2008; Verkuyten & Zaremba, 2005), a stereotype effect was found. That is, this research showed that anger was more quickly perceived for male compared to female faces and for Moroccan-Dutch compared to White-Dutch faces, whereas sadness was more quickly perceived on female compared to male and White-Dutch

☆ This paper has been recommended for acceptance by Dr. Rachael Jack

* Corresponding author.
E-mail address: maximilian.primbs@gmx.de (M.A. Primbs).

[1] These authors contributed equally and are ordered alphabetically.

https://doi.org/10.1016/j.jesp.2022.104394
Received 6 April 2022; Received in revised form 19 July 2022; Accepted 19 July 2022
Available online 2 August 2022

compared to Moroccan-Dutch faces, respectively (= stereotype effects).

On a theoretical level, one framework that can explain these findings are Bayesian models of social perception (Otten, Seth, & Pinto, 2017). Bayesian models, often also referred to as predictive coding frameworks, posit that the perception of a stimulus is influenced by two separate sources of information, namely *the likelihood* function and *the prior*. Consider, for example, a hypothetical experiment in which participants are presented with White-Dutch and Moroccan-Dutch faces displaying either anger or sadness at either 50% or 100% emotion intensity. The likelihood would represent the probability of seeing an emotional expression if there is indeed an emotion displayed or P(visual input | emotion). The likelihood here is influenced by whether the emotion is displayed at 50% or 100% emotion intensity. The prior would represent the probability of seeing an emotional face or P(emotion), which would be influenced by stereotypical associations related to the ethnicity of the face. Finally, the posterior would represent the probability of categorizing a face as angry or sad given the face, or P(emotion | visual input). Following Bayesian models, the perception of the stimulus is influenced by both the prior and the visual input, which means that both sources of information interact: The weaker the visual input (i.e., 50% emotion intensity) and the stronger the prior (i.e., stronger emotion – social category associations), the stronger the tendency to respond in a stereotype-congruent manner regardless of the emotion – social category combination actually being presented.

In the research of Bijlstra et al. (2010), the likelihood can be described as P(face | emotion), the posterior as P(emotion | face), and the prior, which is influenced by participants stereotypical associations between Moroccan-Dutch and anger, and White-Dutch and sadness, respectively, as P(emotion). The experimental design facilitated reliance on the prior, i.e., the stereotypes, because the emotional faces were presented only briefly. Further experiments showed that the strength of the stereotype effect on emotion perception indeed depends on participant's priors. Individuals with stronger stereotype associations showed stronger stereotype effects on the emotion perception task (Bijlstra et al., 2014). In sum, a stronger prior has a larger influence on the subsequent interpretation of the sensory input, and thus leads to a greater stereotype effect, just as generally predicted by predictive coding frameworks.

A further prediction made by predictive coding frameworks is that the influence of the prior should increase as the strength of the visual input decreases (as postulated in our hypothetical example above). Accordingly, behavioral studies have long established that people are more likely to rely on stereotypes and heuristics in ambiguous or uncertain situations (Correll, Hudson, Guillermo, & Ma, 2014; Correll, Park, Judd, & Wittenbrink, 2002; Eberhardt, Goff, Purdie, & Davies, 2004; Neth & Gigerenzer, 2015; Tversky & Kahneman, 1974). Similarly, one study on emotion perception presented participants with racially ambiguous faces and showed that participants more often and faster perceived ambiguous angry faces as Black than as White, i.e., stereotype-congruent (Hutchings & Haddock, 2008).

In the present study, we applied the predictive coding framework (Otten et al., 2017) to the effects of stereotypes on emotion perception (Bijlstra et al., 2010) by testing the prediction that decreasing the strength of the visual input increases reliance on the prior. That means, we expect that decreasing P(face | emotion) or the probability of seeing an emotional expression if there is indeed an emotion displayed should increase the effect of the stereotypical associations between social group and emotion on P(emotion | face) or the probability of categorizing a face as angry or sad.

There are two reasons why this prediction is important to test: Firstly, it is a novel prediction of the predictive coding framework that is not made by dual process models (e.g., Strack & Deutsch, 2004). Secondly, and related, by testing this prediction that is uniquely based on predictive coding framework, it may further help to better understand the applicability of predictive coding frameworks to social cognition in general and emotion perception in particular.

To do so, we relied on a source of uncertainty that was introduced by necessity: Face masks. To limit the spread of the Covid-19 pandemic, many governments recommend the use of surgical face masks and other face coverings to their citizens. Typically, such face masks cover the mouth area and the nose, and often also any remaining area below the eye region. Since the mouth area is fundamental to emotion perception (Blais, Roy, Fiset, Arguin, & Gosselin, 2012), face coverings make it harder to perceive emotions (Carbon, 2020; Rinck, Primbs, Verpaalen, & Bijlstra, 2022). Importantly, perceivers focus on different parts of the face, depending on the emotion displayed. For example, for anger and sadness, perceivers typically focus on the upper part of the face, which means that although there is important information obscured by the mouth masks, perceivers should still be able to perceive the displayed emotion (Smith, Cottrell, Gosselin, & Schyns, 2005). For example, Rinck and colleagues investigated the accuracy of emotion judgements for models displaying the 6 basic emotions and a neutral expression with or without a face mask and showed that face masks drastically reduced the accuracy with which most emotions were perceived, by almost 20% across emotions. In terms of a predictive coding model, this means that face masks decrease the strength of the visual input.

In sum, the present research investigated whether face masks affect the reliance on stereotypical associations between certain social categories and emotions, when perceiving emotions. Across two studies, employing a speeded categorization task (Hugenberg, 2005) and different social categories and stereotypes, participants were asked to categorize emotions when presented with various emotional faces with and without face masks. Study 1 investigated whether face masks increase gender-emotion stereotypes. In Study 2, we extended the findings of Study 1 to ethnicity-emotion stereotypes. Based on predictive coding frameworks, we predicted that decreasing the strength of the visual input will increase reliance on the prior. In short, we expected that introducing face masks increases the size of the stereotype effects.

The present research specification curve analysis demonstrates the use of specification curve analysis (Simonsohn, Simmons, & Nelson, 2020) as a viable and necessary tool for the analysis of reaction time data in social cognition research. Recent surveys of reaction time analyses have shown that researchers vary substantially in the type of reaction time data pre-processing they employ (Kerr, Hesselmann, Räling, Wartenburger, & Sterzer, 2017; Primbs, Holland, Quandt, & Bijlstra, 2022), and that differences in data pre-processing have a considerable influence on the outcomes of statistical tests and potential conclusions that can be drawn from the data (Primbs et al., 2022). Crucially, many of these decisions are arbitrary and there are only few evidence-based guidelines on how to make such decisions available (e.g., André, 2021; Ratcliff, 1993). Recognizing this arbitrariness, specification curve analysis allows researchers to draw inferences and obtain *p*-values across many different data pre-processing and analysis pathways, increasing replicability and decreasing the chance of false positives.

2. Study 1

In Study 1, participants categorized male, and female faces as angry or sad. In line with common gender-emotion stereotypes (Plant et al., 2000), namely that anger is more typical for men and sadness is more typical for women and replicating earlier research on stereotypes and emotion perception (Bijlstra et al., 2010), we expected a two-way interaction between Emotion and Model Gender. That is, we expected that anger is perceived faster on male compared to female faces (male-anger stereotype effect), and that sadness is perceived faster on female compared to male faces (female-sadness stereotype effect). Importantly, and testing the main research question of the present study, we expected that both stereotype effects would be larger for masked compared to unmasked face (i.e., a three-way interaction between Emotion, Model Gender, and Mask Status).

2.1. Methods

2.1.1. Participants

We recruited a sample of 262 adult, English-speaking participants via Prolific. After application of our pre-registered exclusion criteria, a final sample size of 155 participants remained. Please note that most excluded participants ($n = 102$) did not actually complete the experiment – they failed the attention check presented during the instructions and were directly forwarded to the end of the experiment, skipping all experimental trials. The other participants were removed because they were too slow (3*SD* from the mean reaction time; $n = 3$) or made too many mistakes ($n = 2$). The final sample consisted of 97 males and 58 females between 18 and 64 years ($M = 25.58$, $SD = 8.75$) from 25 countries. Participants were paid in accordance with Prolific guidelines on fair compensation and received at least 1.75 British pounds for 14 min of their time. The study was reviewed independently by the Ethics Committee Social Sciences (ECSS) of Radboud University, and there was no formal objection to this study (reference number: ECSW2017–3001-45).

2.1.2. Sensitivity power analysis

We further conducted a sensitivity power analysis for a three-way ANOVA using MorePower (version 6.0.4; Campbell & Thompson, 2012). With $\alpha = .05$, our sample of 155 participants has 80% to detect effect sizes as small as $\eta_p^2 = .049$.

2.1.3. Procedure & materials

Participants signed up on Prolific and were subsequently linked to the Qualtrics platform, where they were presented with an information letter and a consent form. After agreeing to participate, they were instructed to download the Inquisit Web Launcher (2016), which launched the actual experiment. Next, participants filled in demographic information and then completed a speeded categorization task featuring sad and angry faces (Bijlstra et al., 2010). Twenty-four faces (12 male and 12 female actors) were selected from the Radboud Faces Database (Langner et al., 2010), based on average correct emotion identification rates across emotions in the validation study (see Langner et al., 2010). For each actor, the frontal view pictures (90 degree camera angle) displaying the emotions anger and sadness were selected, and a masked version of each image was created by superimposing a surgical face mask, covering the mouth-nose region of the face (see Fig. 1). We used the images actors 01, 02, 03, 04, 07, 08, 12, 14, 20, 22, 23, 24, 25, 27, 28, 31, 32, 33, 46, 47, 49, 57, 58, and 71 of the Radboud Face Database.

The speeded categorization task consisted of 192 trials: 2 Emotion (anger, sadness) * 2 Model Gender (male, female) * 2 Mask Status (masked, unmasked) * 12 Actors, with each unique face being presented twice. For each trial, participants were presented with a fixation cross for 1000 ms, followed by a 280 × 350 pixels large face for 200 ms. Participants were instructed to categorize the emotion displayed on the face as fast and accurately as possible, using the 'a' and the 'l' key (key mapping was counterbalanced across participants). Please note, that participants hereby always had to choose between categorizing a face as sad or as angry. Response key reminders were presented in the left and right top corner of the computer screen. The trials were divided into two blocks of 96 trials and participants had the opportunity to take a break between blocks. The order of trials was fully randomised within blocks and each unique face was presented once per block. Please note that the visual angle – that is the size of the image on the retina calculated from the size of the stimuli and the distance from the stimuli – varied between participants, because they completed the task at home on their personal computers and therefore varied in distance to the stimuli.

After the experiment, participants completed a seriousness check (Aust, Diedenhofen, Ullrich, & Musch, 2013) by indicating either "I have taken part seriously" or "I have just clicked through, please throw my data away" and were informed that their answer to this question would not affect compensation. To further enhance data quality, participants were presented with an attention check as part of the experimental instructions shown at the beginning of the study. The attention check required participants to read the instructions and subsequently do nothing for 20 s, after which they were forwarded to the next page. Participants who failed the attention check were immediately forwarded to the end of the experiment, skipping all experimental trials and thus not contributing any data.

2.1.4. Data analysis

2.1.4.1. Confirmatory analyses. Both reaction times and accuracy data were analysed. Recognizing the large number of possible pre-processing decisions in the analysis of reaction time data, we employed specification curve analysis (Simonsohn et al., 2020).

2.1.4.2. Specification curve analysis. The steps necessary to set up and conduct our specification curve analysis are shown in Fig. 2. First, we considered our design and our independent and dependent variables (*Step 1*), consulted prior research on typical ways to analyse reaction time data obtained (*Step 2*), and we used that information to determine a sensible statistical test – in our case a three-way ANOVA with latency as dependent variable, and Emotion (Sadness vs. Anger), Model Gender (Male vs. Female) and Mask Status (Masked vs. No Mask) as independent variables (*Step 3*). Next, based on our experiences with the analysis of reaction time data and prior research using our paradigm, we determined all sensible pre-processing decisions (*Step 4*) and used those to create a multiverse of sensible pre-processing pathways (*Step 5*). That is, we altered (i) the data transformation employed, (ii) the minimum threshold for reaction times to be included in the analyses, and (iii) the data-based outlier trimming technique. For data transformation, we used either no transformation, log-transformation, or latency-normalisation, which is a procedure that removes between-subject variability in overall reaction times (Gayet & Stein, 2017), resulting in 3 levels. For the minimum threshold, we varied the response time cut-off from 0 ms to 300 ms in steps of 50 ms, resulting in 7 levels. For the data-based outlier trimming method we varied the number of median absolute deviations from the median (Leys, Ley, Klein, Bernard, & Licata, 2013) from 1 to 3 in steps of 0.5 or applied no data-based trimming, resulting in 6 levels. In total, the full combination of data transformation, minimum threshold, and data-based outlier trimming gave rise to a multiverse of 126 data pre-processing pathways (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). Followingly, we considered the equivalence of the different pre-processing pathways (for details please consult Del Giudice & Gangestad, 2021; *Step 6*) and concluded that for our purposes, they can be considered equivalent. We want to explicitly recognize that the specified multiverse represents one of many possible multiverses of analyses. Therefore, we pre-registered the pathways and evaluation criteria used in the multiverse analysis (*Step 7*).

To draw statistical inferences from this multiverse, we first conducted 126 within-subjects ANOVAs; one for each of the 126 possible pre-processing pathways in the multiverse (*Step 8*). Subsequently, we calculated the median effect size across all pathways and the number of significant pathways for each effect of interest respectively. Then, to obtain *p*-values, we compared these test statistics to test statistics from datasets where the null hypothesis is true.[2] To obtain such datasets, we employed a resampling technique called permutation testing using a four-step approach (*Step 9*; Simonsohn et al., 2020): First, we randomly shuffled all independent variables. By randomly shuffling the independent variables, we created a dataset where the null hypothesis is true by construction, but which maintains all other properties of the observed

[2] Note: This is necessary because the different pre-processing pathways are correlated with another and violate the NHST assumption that datapoints are independent.

Fig. 1. Example of an unmasked and masked sad female face. Faces were presented.

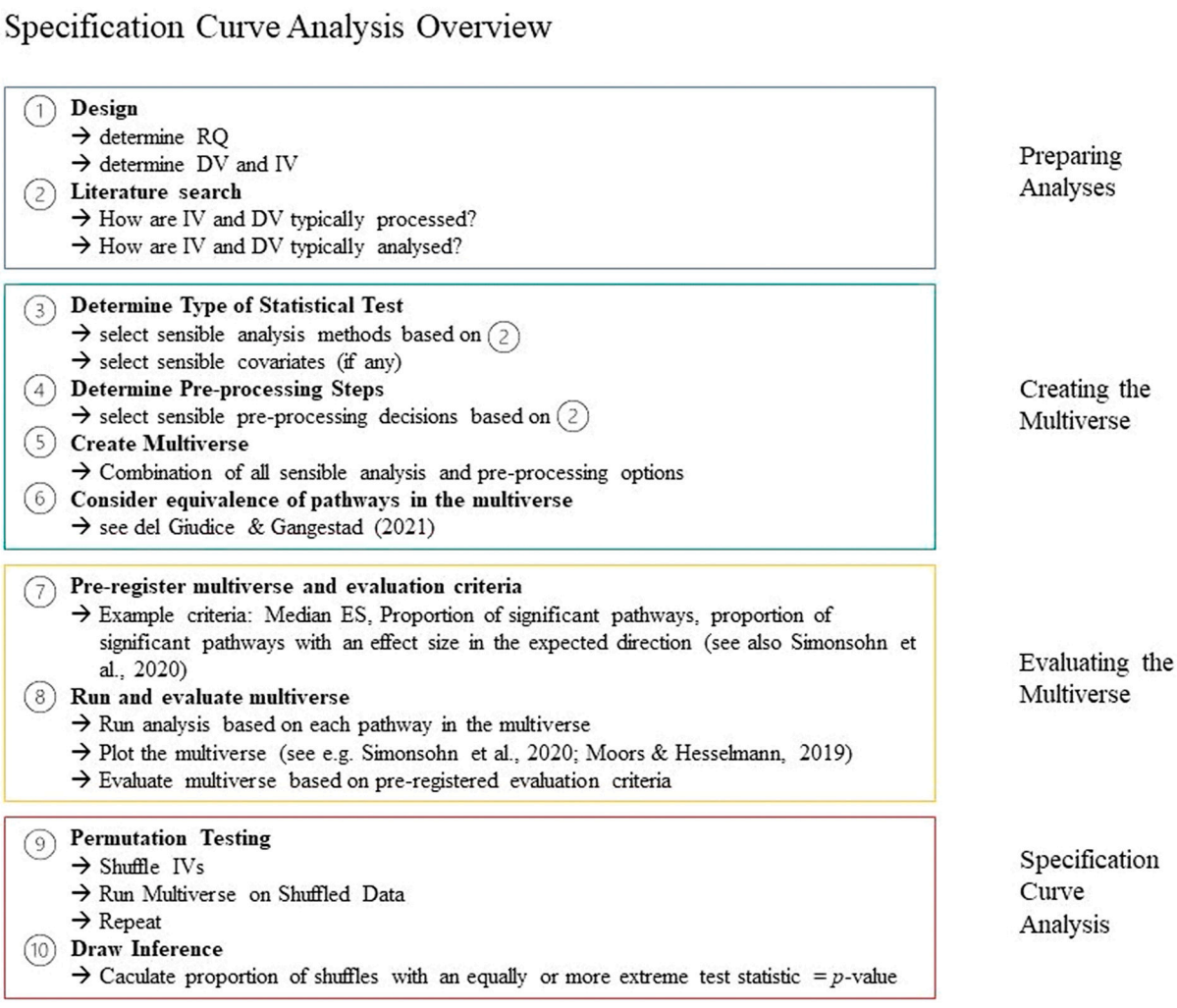

Fig. 2. Overview chart of all steps required to conduct a specification curve analysis. For users who wish to merely conduct a multiverse analysis, steps 9 and 10 can be skipped.

data, such as within-subjects or between pre-processing pathway correlations. Second, we re-ran the 126 ANOVAs based on the multiverse of pre-processing pathways. Third, we repeated this process of shuffling and re-running the analyses 500 times. Finally, we compared the median effect size and the number of significant pathways found in the observed data with the same test statistics of the permutated datasets. That is, we calculated the proportion of permutations which yielded an equally large or larger median effect size and the same or a larger number of significant pathways than the observed data. This proportion of test statistics obtained under the null hypothesis corresponds to a classical *p*-

value (*Step 10*; Simonsohn et al., 2020). A p-value of <.002 thus means that not a single permutation yielded equally or more extreme test statistics than the observed data.

We followed up on the ANOVA by conducting follow-up *t*-tests aimed at the focal contrasts. First, to further scrutinize the Emotion * Model Gender interaction, we conducted multiverse *t*-tests investigating whether the difference between male and female faces is significant for both anger and sadness. The *t*-tests and the subsequent permutation tests followed the same procedure as the ANOVA described above, with one crucial difference: Cohen's *d* can distinguish directional hypotheses, whereas η_p^2 cannot, and thus we also calculated the number of significant pre-processing pathways with an effect in the expected direction as third test statistic.

Second, to further investigate the Emotion * Model Gender * Mask Status interaction, we conducted multiverse *t*-tests to test whether the effect of masks on the difference scores between male and female faces is significant and in the respective expected direction for both anger and sadness. Hereby we followed the same procedure as for the Emotion * Model Gender interaction.

Third, to see whether we replicated the results of Bijlstra et al. (2010), we investigated the Emotion * Model Gender interaction in the subset of unmasked faces. Here we used multiverse *t*-tests and permutation testing identical to the tests described for the Emotion * Model Gender interaction in the full sample.

2.1.4.3. Accuracy analyses. For the accuracy data, we conducted a 3-way within-subjects ANOVA with proportion of correct responses as dependent variable, and Emotion (Sadness vs. Anger), Model Gender (Male vs. Female) and Mask Status (Masked vs. No Mask) as independent variables. To further investigate the accuracy data, we conducted follow-up *t*-tests for each significant interaction. All accuracy analyses were completed on the dataset with no transformation or outlier removal applied. We define accuracy as the proportion of responses in which a participant perceives the emotion as the actor intended to display it (see Langner et al., 2010). We recognize that the facial expression displayed by the actor does not relate to the internal state of the actor (see Feldman Barrett, Adolphs, Marsella, Martinzed, & Pollak, 2019).

2.1.4.4. Exploratory analyses. To further explore our data, we tested the possibility that face masks introduced a stereotype-congruent response bias. To that end, we calculated hits (defined as angry face / angry response), false alarms (sad face / angry response), misses (angry face / sad response) and correct rejections (sad face / sad response) for each participant.[3] Afterwards, we calculated the sensitivity measure *d'* (d prime) and the response bias indicator beta (Pallier, 2002). For each of the two measures, we conducted a 2-way within-subjects ANOVA with the respective measure as dependent variable, and Model Gender (Male vs. Female) and Mask Status (Masked vs. No Mask) as independent variables.

Moreover, to fully explore the Model Gender by Emotion interaction, we also investigated the within-gender contrasts. Further explored the Model Gender by Emotion interaction by investigating within-gender contrasts. That is, we conducted a specification curve analysis identical to the main analyses described above based on multiverse *t*-tests with reaction time as dependent variable and emotion as independent variable for each gender separately.

2.1.5. Transparency statement

All confirmatory analyses were pre-registered on the Open Science Framework (https://osf.io/pz4xh/?view_only=c05a128203944b9e81fdf7e285ef909e), whereas all exploratory analyses were conducted post-hoc with the aim of further understanding the data. The confirmatory analyses did not deviate from the pre-registered analyses. The data and analysis code are accessible on the Open Science Framework. The functions used to run the multiverse analysis were programmed by the authors and the respective version of the functions used in each study is also available on the OSF. An overview of all packages we used can be found in Appendix A. We report all measures, manipulations, and exclusions used in Study 1.

2.2. Results

2.2.1. Confirmatory analyses

2.2.1.1. Reaction times. The multiverse ANOVA and the subsequent specification curve analysis of the reaction time data showed that the number of significant specifications for the Emotion * Model Gender interaction (126/126 specifications, $p = .002$)[4] and the significant Emotion * Model Gender * Mask Status interaction (84/126 specifications, $p = .026$) was significantly higher than would be expected if the null hypothesis were true. Further investigation of the Emotion * Model Gender interaction revealed that only the male-anger stereotype (126/126 specifications, $p = .004$) – but not the female-sadness stereotype (0/126 specifications, $p = 1$) – was observed in the data. That means, anger was perceived more quickly on male faces compared to female faces, but sadness was not perceived more quickly on female faces compared to male faces (Table 2).

For the Emotion * Model Gender * Mask Status interaction, follow-up *t*-tests showed that the effect of mask on the stereotype effects was also only significant for the male-anger stereotype (84/126 specifications, $p = .016$), but not for the female-sadness stereotype (0/126 specifications, $p = 1$). Importantly, whereas we hypothesized that the stereotype effects would be *larger* for masked than unmasked faces, the data indicated that the male-anger stereotype effect was *smaller* in masked than unmasked faces (84/126 specifications, $p = .022$. Table 1 and Fig. 3 provide an overview over the full test results. Table 2 displays the summary statistics.

2.2.1.2. Accuracy. The Emotion * Model Gender * Mask Status ANOVA on the error rates showed significant main effects of Emotion, $F(1, 154) = 5.038$, $p = .026$, $\eta_p^2 = .031$, Model Gender, $F(1, 154) = 5.617$, $p = .019$, $\eta_p^2 = .035$, and Mask Status, $F(1, 154) = 14.031$, $p < .001$, $\eta_p^2 = .084$. That is, sad faces ($M = 0.818$, $SD = 0.386$) were perceived more accurately than angry faces ($M = 0.802$, $SD = 0.398$), male faces ($M = 0.816$, $SD = 0.388$) were perceived more accurately than female faces ($M = 0.805$, $SD = 0.396$), and unmasked faces ($M = 0.820$, $SD = 0.385$) were perceived more accurately than masked faces ($M = 0.801$, $SD = 0.400$). There was a significant Mask Status * Emotion interaction, $F(1, 154) = 56.981$, $p < .001$, $\eta_p^2 = 0.270$. Follow-up *t*-tests revealed that sad masked faces were indeed perceived with lower accuracy than sad unmasked faces, $t(154) = 7.829$, $p < .001$, Cohen's $d = 0.661$. In contrast, angry masked faces were perceived with *higher* accuracy than angry unmasked faces, $t(154) = 3.873$, $p < .001$, Cohen's $d = 0.265$. The Emotion * Model Gender interaction, $F(1, 154) = 1.864$, $p = .174$, $\eta_p^2 = 0.012$, the Model Gender * Mask Status interaction, $F(1, 154) = 2.177$, $p = .142$, $\eta_p^2 = .014$, and the Emotion * Model Gender * Mask Status interaction, $F(1, 154) = 1.106$, $p = .294$, $\eta_p^2 = .007$, were non-significant.

[3] Please note that defining hits as correct identification of anger is arbitrary: We could have also defined hits as correct identification of sadness. This decision does not influence the interpretation of the results.

[4] The *p*-values reported in the text are based on the number of significant specifications. The *p*-values corresponding to the other pre-registered test statistics are reported in the Tables 1 and 3.

Table 1.
Overview specification curve analysis.

Effect of Interest	Test Statistic	Observed Result	P-value (% of shuffled sample with results as, or more extreme)
ANOVA			
Emotion*Model Gender	(1) Median effect size	$\eta_p^2 = .099$	***p* < .002**
	(2) Number significant	126/126	***p* = .002**
Emotion*Model Gender * Mask Status	(1) Median effect size	$\eta_p^2 = .038$	***p* = .006**
	(2) Number significant	84/126	***p* = .026**
T-Tests Emotion*Model Gender Follow-up: Within Emotion Contrast			
Angry Subset	(1) Median effect size	$d = 0.253$	***p* < .002**
	(2) Number significant	126/126	***p* = .004**
	(3) Number significant in expected direction	126/126	***p* = .004**
Sad Subset	(1) Median effect size	$d = 0.012$	$p = .706$
	(2) Number significant	0/126	$p = 1$
	(3) Number significant in expected direction	0/126	$p = .844$
T-Tests Emotion*Model Gender * Mask Status: Effect of Mask on Within Emotion Contrast			
Angry Subset	(1) Median effect size	$d = 0.282$	***p* = .016**
	(2) Number significant	84/126	***p* = .022**
	(3) Number significant in expected direction	0/126	$p = .860$
	(4) Number significant in direction of observed effect	84/126	***p* = .002**
Sad Subset	(1) Median effect size	$d = 0.081$	$p = .430$
	(2) Number significant	0/126	$p = 1$
	(3) Number significant in expected direction	0/126	$p = .810$
T-Tests Emotion* Model Gender Unmasked Subset: Within Emotion Contrast			
Angry Subset	(1) Median effect size	$d = 0.319$	***p* < .002**
	(2) Number significant	126/126	***p* < .002**
	(3) Number significant in expected direction	126/126	***p* < .002**
Sad Subset	(1) Median effect size	$d = 0.033$	$p = .466$
	(2) Number significant	21/126	$p = .156$
	(3) Number significant in expected direction	21/126	$p = .144$

2.2.2. Exploratory analyses

2.2.2.1. Response bias. The Model Gender * Mask Status ANOVA on the *d*' scores showed significant effects of Model Gender, $F(1, 153) = 6.24$, $p = .014$, $\eta_p^2 = 0.039$, and Mask Status, $F(1, 153) = 12.44$, $p < .001$, $\eta_p^2 = .075$. Emotions on unmasked faces ($M = 0.551$, $SD = 0.157$) were perceived more accurately on masked faces ($M = 0.522$ $SD = 0.167$) and more accurately on male faces ($M = 0.546$, $SD = 0.162$) than on female faces ($M = 0.527$, $SD = 0.163$). The Model Gender * Mask Status interaction effect was non-significant, $F(1,153) = 1.81$, $p = .181$, $\eta_p^2 = .013$. The same Model Gender * Mask Status ANOVA on the beta scores showed a significant effect of face mask, $F(1, 153) = 42.80$, $p < .001$, $\eta_p^2 = .219$, indicating a response bias towards anger in masked faces ($M = 0.997$, $SD = 0.033$) compared to unmasked faces ($M = 1.015$, $SD = 0.032$, Cohen's $d = 0.551$). The main effect of Model Gender, $F(1, 153) = 0.00$, $p = .979$, $\eta_p^2 < .001$, and the interaction, $F(1, 153) = 0.06$, $p = .813$, $\eta_p^2 < .001$, were non-significant.

2.2.2.2. Within-gender contrasts. Finally, we conducted a specification curve analysis identical to the main analyses described above based on multiverse *t*-tests with reaction time as dependent variable and emotion as independent variable for each gender separately. The multiverse *t*-tests and subsequent specification curve analysis for the effect of emotion in the subset of male faces revealed that the median effect size ($d = 0.487$, $p < .002$) the number of significant specifications (126/126, $p < .002$), and the number of significant specifications in the observed direction (126/126, $p < .002$) were all significantly higher than would be expected if the null hypothesis were true. Likewise, the analysis of the female faces showed that the median effect size ($d = 0.21$, $p < .002$), the number of significant specifications (126/126, $p = .002$), and the number of significant specifications in the observed direction (126/126, $p = .002$) were all significantly higher than would be expected if the null hypothesis were true. That means that for both male and female faces, anger is perceived faster than sadness.

2.3. Discussion

The goal of Study 1 was to investigate whether face masks affect the reliance on stereotypical associations of male and female faces with anger and sadness. In line with our predictions, we successfully replicated the male-anger stereotype effect (Bijlstra et al., 2010), with anger being perceived more quickly on male compared to female faces. However, in contrast to our predictions, we found that face masks *decreased* the size of this male-anger stereotype effect. Moreover, we failed to find a female-sadness stereotype effect, with sadness not being perceived more quickly on female compared to male faces, and consequently we did not find an effect of face mask on sad faces. Notably, following Bayesian models of social cognition (Otten et al., 2017), the observed direction of the effect of face masks on angry faces is very unlikely, and presents a challenge to the suitability of Bayesian models for explaining stereotype effects in speeded categorization paradigms.

3. Study 2

Study 1 provided inconclusive evidence about the effect of face masks on the stereotype effect in emotion perception. Thus, with Study 2 we conceptually replicated Study 1 by extending our investigation to ethnicity-emotion stereotype associations. In line with societal stereotypes that anger is more typical of Moroccan-Dutch males than of White-Dutch males, we expected anger to be perceived faster for Moroccan-Dutch compared to White-Dutch faces (Moroccan-anger stereotype effect). Moreover, although neither sadness nor anger are more typical of Dutch faces in general, compared to Moroccan-Dutch faces we would expect sadness to be perceived faster for White-Dutch compared to Moroccan-Dutch faces (Bijlstra et al., 2014). Following a predictive coding framework, we would expect again that these effects are larger for masked compared to unmasked faces. In contrast, following the results of Study 1, we would expect that these effects are smaller for masked faces compared to unmasked faces.

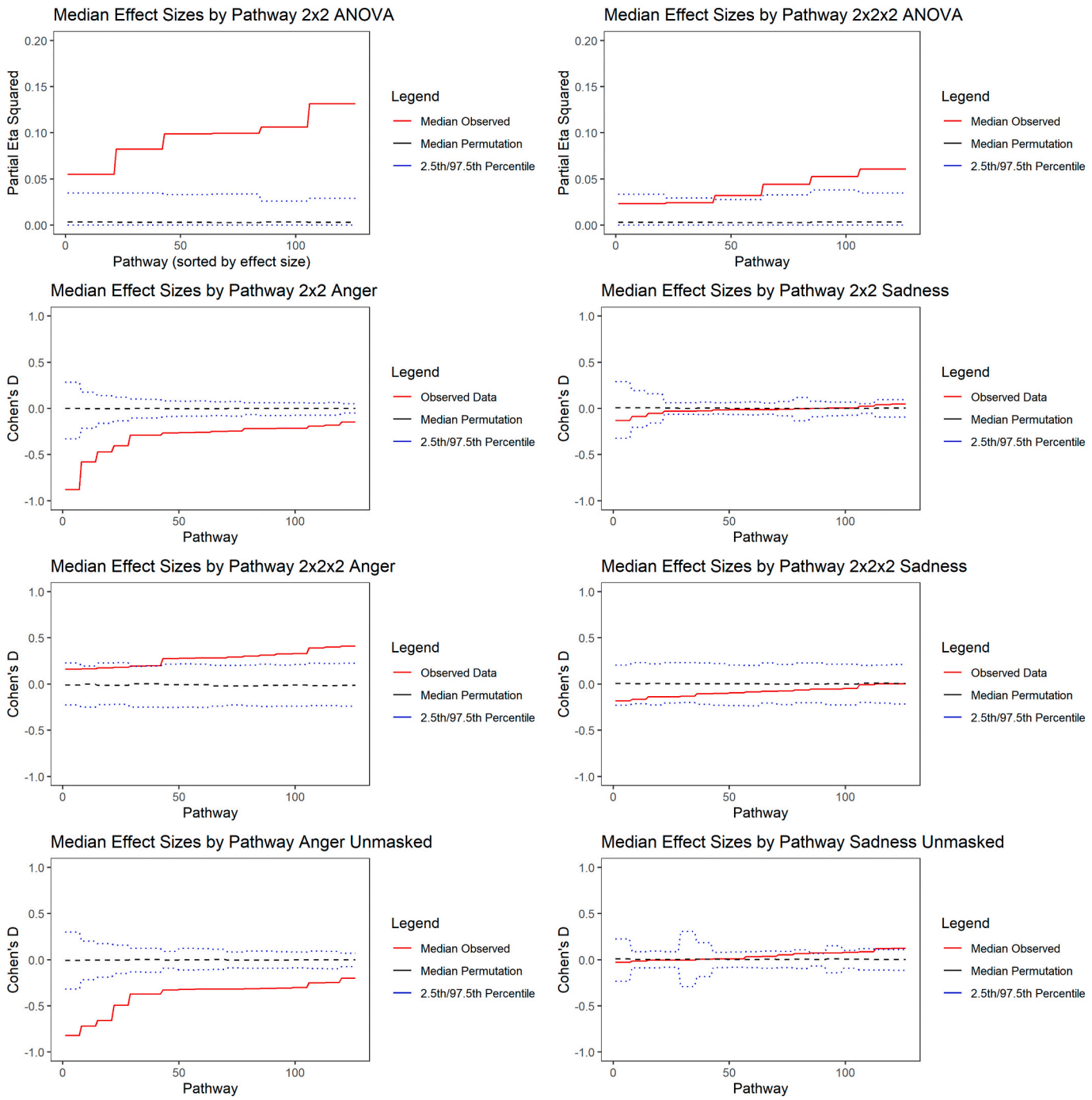

Fig. 3. Top row shows the results of the specification curve analysis for the ANOVA. Second row from the top shows the results of the specification curve analysis of the *t*-tests following up on the Emotion * Model Gender interaction separated by emotion. Third row from the top shows the results of the specification curve analysis of the *t*-tests following up on the Emotion * Model Gender * Mask Status interaction separated by emotion. Bottom row shows the results of the specification curve analysis following up on the Emotion * Model Gender interaction in the unmasked subset, separated by Emotion. The red line shows the median effect size of the observed data by specification. The dashed line shows the median effect size of the 500 permutation tests and thus provides visual confirmation that the shuffling indeed produced datasets where the null hypothesis is true – i.e., datasets with an effect size of roughly zero. The dotted lines show the 2.5th and 97.5th percentile of the median effect size of the permutation tests. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1. Participants

We recruited a sample of 203 adult, White, and English-speaking participants via Prolific. To increase the likelihood that participants were exposed to societal stereotypes related to Moroccan males, only participants residing in Germany, the Netherlands, France, or Belgium were eligible to participate. We applied our pre-registered exclusion criteria and excluded participants who did not complete the whole experiment ($n = 19$), who indicated that they did not participate seriously (n = 1), who made too many errors (3*SD* from the mean; $n = 4$) or who were too slow (3*SD* from the mean reaction time; n = 4). The final sample of 175 participants consisted of 97 males and 78 females between 18 and 61 years ($M = 29.49$, $SD = 9.08$) of 28 nationalities. Participants were paid in accordance with Prolific guidelines on fair

Table 2
Overview of raw reaction times and accuracy data without any transformations or trimming applied (Study 1). Values in latency columns reflect milliseconds. Values in accuracy columns reflect percent correct.

Emotion	ModelGender	M(SD) latency masked	M(SD) latency unmasked	M(SD) accuracy masked	M(SD) accuracy unmasked
Sad	Female	773 (219)	718 (207)	0.786 (0.411)	0.850 (0.357)
Sad	Male	768 (224)	718 (198)	0.786 (0.411)	0.856 (0.351)
Angry	Female	721 (201)	737 (218)	0.812 (0.391)	0.771 (0.420
Angry	Male	710 (203)	708(217)	0.819 (0.386)	0.800 (0.400)

compensation and received at least 1.75 British pounds for 14 min of their time.

3.2. Sensitivity power analysis

We further conducted a sensitivity power analysis using MorePower (version 6.0.4; Campbell & Thompson, 2012). With $\alpha = .05$, our sample of 175 participants has 80% to detect effect sizes as small as $\eta_p^2 = .044$.

3.3. Procedure & materials

The procedure was mostly identical to Study 1, with two notable differences. First, as Study 2 focused on ethnicity, we replaced the sample of female faces from the Radboud Face Database (Langner et al., 2010) with a sample of male Moroccan-Dutch faces from the Radboud Faces Database. We used the images of actors 03, 07, 20, 23, 24, 25, 28, 29, 33, 45, 46, 47, 49, 50, 51, 52, 55, 59, 67, 69, 70, 71, 72, and 73. Second, in line with Prolific policy, we changed the attention check of Study 1, and we now instructed participants to select a specific answer on two questions ostensibly related to the research at hand.

3.4. Data analysis

3.4.1. Confirmatory analyses

The confirmatory analyses were identical to the confirmatory analyses of Study 1 (see https://osf.io/6ayd5/?view_only=bce1ec71085d40aeac7963ff7cf77b73). However, we investigated the effects of model ethnicity instead of the effects of model gender.

3.4.2. Exploratory analyses

The exploratory analyses were identical to the exploratory analyses of Study 1. However, we again focussed on model ethnicity instead of model gender.

3.5. Transparency statement

As in Study 1, all hypotheses and analyses were pre-registered and data, analysis scripts and functions employed to implement the specification curve analysis are available on the OSF (https://osf.io/vcwbu/files?view_only=eb2f0ee86614481eb7f354168963a40e). We report all measures, manipulations, and exclusions used in Study 2.

3.6. Results

3.6.1. Reaction times

The multiverse ANOVA, and subsequent specification curve analysis on the reaction time data, showed that the number of significant specifications for the Emotion * Model Ethnicity interaction was significantly higher than would be expected if the null hypothesis were true (126/126 specifications, $p < .002$). Follow-up tests showed that anger was perceived faster on Moroccan-Dutch compared to White-Dutch faces (126/126, $p < .002$) and sadness was perceived faster on White-Dutch compared to Moroccan-Dutch faces ((126/126, $p < .002$; see Table 4). For the Emotion * Model Ethnicity * Mask Status interaction, the number of significant interactions was not significantly higher than would be expected if the null hypothesis were true (0/126, $p = 1$). Follow-up tests indicate that indeed neither the association between Dutch-sadness (21/126, $p = .058$) nor the Moroccan-angry stereotype (0/126, $p = .784$) were influenced by face masks. Notably, anger was detected both faster

Table 3
Overview of specification curve analysis.

Effect of interest	Test statistic	Observed result	P-value (% of shuffled sample with results as, or more extreme)
ANOVA			
Emotion*Model Ethnicity	(1) Median effect size	$\eta_p^2 = .286$	**$p < .002$**
	(2) Number significant	126/126	**$p < .002$**
Emotion*Model Ethnicity * Mask Status	(1) Median effect size	$\eta_p^2 = .010$	$p = .134$
	(2) Number significant	0/126	$p = 1$
T-Tests Emotion*Model Ethnicity: Within Emotion Contrast			
Angry Subset	(1) Median effect size	$d = 0.298$	**$p < .002$**
	(2) Number significant	126/126	**$p = .002$**
	(3) Number significant in expected direction	126/126	**$p < .002$**
Sad Subset	(1) Median effect size	$d = 0.155$	**$p < .002$**
	(2) Number significant	126/126	**$p < .002$**
	(3) Number significant in expected direction	126/126	**$p < .002$**
T-Tests Emotion*Model Ethnicity * Mask Status: Effect of Mask on Within Emotion Contrast			
Angry Subset	(1) Median effect size	$d = 0.037$	$p = .694$
	(2) Number significant	0/126	$p = 1$
	(3) Number significant in expected direction	0/126	$p = .784$
Sad Subset	(1) Median effect size	$d = 0.149$	$p = .112$
	(2) Number significant	21/126	$p = .130$
	(3) Number significant in expected direction	21/126	$p = .058$
T-Tests Emotion* Model Ethnicity Unmasked Subset: Within Emotion Contrast			
Angry Subset	(1) Median effect size	$d = 0.267$	**$p < .002$**
	(2) Number significant	126/126	**$p < .002$**
	(3) Number significant in expected direction	126/126	**$p < .002$**
Sad Subset	(1) Median effect size	$d = 0.125$	**$p < .002$**
	(2) Number significant	105/126	**$p = .01$**
	(3) Number significant in expected direction	105/126	**$p = .01$**

and more accurately on Moroccan-Dutch compared to White-Dutch faces, and on masked compared to unmasked faces. Table 3 and Fig. 4 provide an overview over the full test results.

3.6.2. Accuracy

The Emotion * Model Ethnicity * Mask Status ANOVA of the error rates showed significant main effects of Emotion, $F(1, 174) = 31.418$, $p < .001$, $\eta_p^2 = .153$, Model Ethnicity, $F(1, 174) = 107.647$, $p < .001$, $\eta_p^2 = .382$, and Mask Status, $F(1, 174) = 5.690$, $p = .018$, $\eta_p^2 = .082$. That is, sad faces ($M = 0.860$, $SD = 0.347$) were perceived more accurately than angry faces ($M = 0.803$, $SD = 0.398$), Moroccan-Dutch faces ($M = 0.853$, $SD = 0.355$) were perceived more accurately than White-Dutch faces ($M = 0.810$, $SD = 0.393$), and unmasked faces ($M = 0.836$, $SD = 0.370$) were perceived more accurately than masked faces ($M = 0.826$, $SD =$

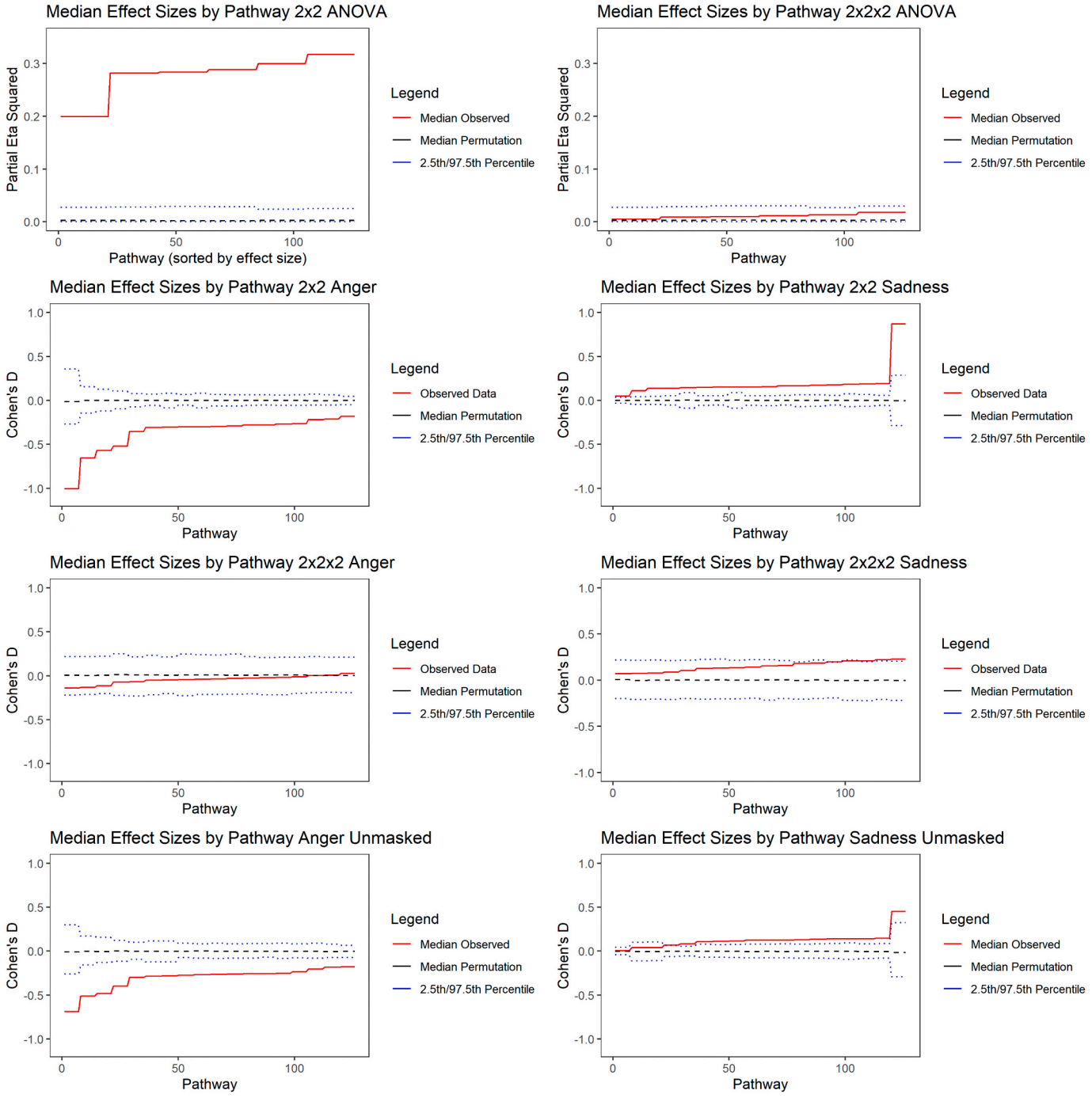

Fig. 4. Top row shows the results of the specification curve analysis for the ANOVA. Second row from the top shows the results of the specification curve analysis of the *t*-tests following up on the Emotion * Model Ethnicity interaction separated by emotion. Third row from the top shows the results of the specification curve analysis of the *t*-tests following up on the Emotion * Model Ethnicity * Mask Status interaction separated by emotion. Bottom row shows the results of the specification curve analysis following up on the Emotion * Model Ethnicity interaction in the unmasked subset, separated by Emotion. The red line shows the median effect size of the observed data by specification. The dashed line shows the median effect size of the 500 permutation tests and thus provides visual confirmation that the shuffling indeed produced datasets where the null hypothesis is true – i.e., datasets with an effect size of roughly zero. The dotted lines show the 2.5th and 97.5th percentile of the median effect size of the permutation tests. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4.
Overview of raw reaction times and accuracy data without any transformation or trimming applied (Study 2). Values in latency columns reflect milliseconds. Values in accuracy columns reflect percent correct.

Emotion	Model Ethnicity	M(SD) latency masked	M(SD) latency unmasked	M(SD) accuracy masked	M(SD) accuracy unmasked
Sad	White-Dutch	740(212)	691(193)	0.848 (0.359)	0.893 (0.310)
Sad	Moroccan-Dutch	770 (226)	706.(208)	0.825 (0.380)	0.873 (0.333)
Angry	White-Dutch	725 (200)	730(218)	0.765 (0.424)	0.732 (0.443)
Angry	Moroccan-Dutch	697 (202)	705 (217)	0.865 (0.342)	0.847 (0.360)

0.379). There also was a significant Emotion * Model Ethnicity interaction, $F(1, 174) = 202.577$, $p < .001$, $\eta_p^2 = 0.538$, and a significant Mask Status * Emotion interaction, $F(1, 157) = 38.876$, $p < .001$, $\eta_p^2 = 0.183$. Follow-up *t*-tests revealed that as expected, sad masked faces were perceived with lower accuracy compared to sad unmasked faces, $t(174) = -6.475$, $p < .001$, Cohen's $d = 0.467$. In contrast, angry masked faces were perceived with *higher* accuracy than angry unmasked faces, $t(174) = 3.558$, $p < .001$, Cohen's $d = 0.197$. Finally, Moroccan-Dutch angry faces were perceived with greater accuracy than White-Dutch angry faces, $t(174) = 16.452$, $p < .001$, Cohen's $d = 0.835$, whereas White-Dutch sad faces were perceived with greater accuracy than Moroccan-Dutch sad faces, $t(174) = -3.780$, $p < .001$, Cohen's $d = 0.230$. The Model Ethnicity * Mask Status interaction, $F(1, 174) = 2.085$, $p = .151$, $\eta_p^2 = .012$, and the Emotion * Model Ethnicity * Mask Status interaction, $F(1, 174) = 0.657$, $p = .419$, $\eta_p^2 = .004$, were non-significant.

3.6.3. Exploratory analyses

3.6.3.1. Response bias. The Model Ethnicity* Mask Status ANOVA of the *d*' scores showed significant main effects of Model Ethnicity, $F(1, 174) = 67.58$, $p < .001$, $\eta_p^2 = .280$, and Mask Status, $F(1, 174) = 9.69$, $p = .002$, $\eta_p^2 = .053$. Emotions on unmasked faces ($M = 0.571$, $SD = 0.147$) were perceived with higher accuracy than emotions on masked faces ($M = 0.547$, $SD = 0.160$), and emotions on Moroccan-Dutch faces ($M = 0.588$, $SD = 0.154$) were perceived with higher accuracy than on White-Dutch faces ($M = 0.529$, $SD = 0.148$). The Model Ethnicity * Mask Status interaction effect was non-significant, $F(1,174) = 0.74$, $p = .392$, $\eta_p^2 = .004$. The Model Ethnicity * Mask Status ANOVA of the beta scores showed a significant effect of Mask Status, $F(1, 174) = 10.41$, $p = .001$, $\eta_p^2 = .056$, indicating a stronger response bias towards anger in masked faces ($M = 0.993$, $SD = 0.094$) than in unmasked faces ($M = 1.000$ $SD = 0.0989$, Cohen's $d = 0.079$). The main effect of Model Ethnicity, $F(1, 174) = 36.72$, $p < .001$, $\eta_p^2 = .174$, was significant, indicating a stronger response bias towards anger in Moroccan-Dutch faces ($M = 0.988$, $SD = 0.097$) than in White-Dutch faces ($M = 1.004$, $SD = 0.095$, Cohen's $d = 0.166$). The Model Ethnicity * Mask Status interaction, $F(1, 174) = 3.81$, $p = .053$, $\eta_p^2 = .021$, was non-significant.

3.6.3.2. Within-ethnicity contrasts. Finally, we conducted a specification curve analysis identical to the main analyses described above based on multiverse *t*-tests with reaction time as dependent variable and emotion as independent variable for each ethnicity separately. The multiverse *t*-tests and subsequent specification curve analysis for the effect of emotion in the subset of Moroccan-Dutch faces revealed that the median effect size ($d = 0.397$, $p < .002$) the number of significant specifications (126/126, $p < .002$), and the number of significant specifications in the observed direction (126/126, $p < .002$) were all significantly higher than would be expected if the null hypothesis were true. The analysis of the White-Dutch faces showed that the median effect size ($d = -0.072$, $p = .022$), and the number of significant specifications in the expected direction (49/126, $p = .022$) were all significantly higher than would be expected if the null hypothesis was true, whereas the absolute number of significant specifications (49/126, $p = .052$) was not higher than would be expected if the null hypothesis was true. These exploratory analyses support the conclusions of the main analyses, namely that there is evidence for a Moroccan-anger stereotype, and weaker but still significant evidence for the idea that sadness is more strongly associated with White-Dutch compared to Moroccan-Dutch faces.

3.7. Discussion

The goal of Study 2 was to investigate whether face masks affect the reliance on stereotypical associations between Moroccan-Dutch and White-Dutch faces and anger and sadness. Firstly, we predicted that anger would be perceived faster on Moroccan-Dutch compared to White-Dutch faces (Moroccan-anger stereotype), and that sadness would be perceived faster on White-Dutch compared to Moroccan-Dutch faces. Notably, we found evidence for both hypotheses, and thus successfully replicated prior studies (Bijlstra et al., 2010, 2014). Secondly, we compared predictions derived from predictive coding accounts (Otten et al., 2017) with predictions derived from Study 1 and found evidence for neither.

4. General discussion

In the current studies, we replicated existing research showing stereotype effects in emotion perception (Bijlstra et al., 2010, 2014). More precisely, we successfully replicated Male-anger and Moroccan-anger stereotypes, whereas we found no evidence for a Female-sadness stereotype. In addition, we showed that sadness is more strongly associated with White-Dutch compared to Moroccan-Dutch faces. These findings are well in line with predictive coding frameworks: The stereotypes influence the prior and interact with the emotional faces to influence the speed and accuracy the emotions are perceived with.

Notably, the effect sizes for the anger-related stereotypes were considerably larger than the effect sizes for sadness-related stereotypes. Following theoretical frameworks that argue that stereotypes are culturally ingrained knowledge (Devine, 1989; Payne, Vuletich, & Lundberg, 2017), this hints at a larger prevalence or greater importance of anger-related stereotypes in the contemporary cultural environment. Alternatively, one may argue that anger-related stimuli signal threat and are thus detected faster by specialised threat detection pathways (Tamietto & de Gelder, 2010). For example, earlier research showed that people detect animals associated with threat faster than other animals (Öhman, Flykt, & Esteves, 2001) and find threatening faces in a crowd more easily than non-threatening faces (Öhman, Lundqvist, & Esteves, 2001). Likewise, seeing Moroccan-Dutch faces may activate danger-related stereotypes and thus facilitate processing more strongly than White-Dutch faces for whom there are weaker danger-related stereotypes. Both explanations are congruent with predictive coding frameworks, as they simply describe different ways in which the prior formation may have taken place.

However, the main research question investigated in the present paper was whether decreasing the strength of the visual input increases reliance on the prior. Across two studies using face masks, gender-emotion, and ethnicity-emotion stereotypes we did not find evidence for this prediction. Importantly, predictive coding frameworks can neither explain the absence of the effect in Study 2 nor its reversal in Study 1. Theoretically, decreasing the probability of seeing an emotional expression if there is indeed an emotion displayed should increase the effect of the stereotypical associations between social group and emotion on the probability of categorizing a face as sad or angry. Even so, whereas increased reaction times for masked faces show that we indeed decreased the strength of the visual input, the relationship between prior and visual input is not necessarily linear. That means, that decreasing the strength of the visual input by some unit need not always

lead to an increase in reliance on the prior by some unit. As such, the absence of the hypothesized effects could be due to a lack of potency of our manipulation. That is, the face masks may not have impaired emotion perception enough to increase reliance on prior stereotypes in a detectable way (Blais et al., 2012). However, in the supplementary materials we present two additional studies which degrade the strength of the visual input by means of noise patterns superimposed on the faces. These studies reveal similar outcomes and replicate earlier stereotype effects (Bijlstra et al., 2010), but again, these were not qualified by the strength of the visual input and thus provide further evidence not in line with predictive coding frameworks (Supplementary Materials S1). Importantly, an explanation arguing that our face mask manipulation was insufficient cannot account for the findings of Study 1, which showed effects opposite to those predicted by predictive coding accounts. Together, we argue that it is unlikely that the absence of the effect is caused by our manipulation.

Dual-process models on the other hand provide a possible explanation for this finding (Strack & Deutsch, 2004): More ambiguous stimuli need to be processed more deliberatively, decreasing the potential effects of stereotypes. As masked faces were generally processed more slowly, they might have been processed more deliberately, potentially decreasing the size of the stereotype effect. In line with this hypothesis research on multiracial faces has shown that people are slower in categorizing multiracial (i.e., more ambiguous) faces compared to monoracial faces and that cognitive load interferes with the categorization of multiracial but not monoracial faces (Chen & Hamilton, 2012). Moreover, the response bias towards anger observed in masked faces may have potentially further reduced the size of the stereotype effects. That means, face masks in themselves might trigger particular responses when categorizing emotions.

In addition to our theoretical goals of understanding stereotype effects in social perception, we also demonstrated the value of specification curve analysis (Simonsohn et al., 2020) as an analysis tool in social cognition research. That is, in both Studies 1 and 2 some effects were only present in a subset of the multiverse of pre-processing pathways. Hence, researchers choosing any single pathway may erroneously conclude the presence or absence of a particular effect. For example, when looking at the categorization advantage for White-Dutch sad faces over Moroccan-Dutch sad faces, we find that some pre-processing pathways were significant (21/126), indicating differences between pre-processing choices. As such, analyses based on any of these 21 pathways would have concluded that face masks increase reliance on these stereotype associations. However, for the remaining 105 pathways, no such conclusion would have been justifiable. Yet, by conducting analyses across pre-processing pathways and comparing test statistics of the observed data with test statistics of datasets where the null hypothesis is true, we found that the number of significant pre-processing pathways is not different from what would be expected if the null hypothesis were true. That means, specification curve analysis has the potential to reduce the proportion of false positive findings in the scientific literature and shorten discussions about optimal analyses strategies (e.g., Christensen & Christensen, 2014; Jung, Shavitt, Viswanathan, & Hilbe, 2014a, 2014b; Malter, 2014; Munoz & Young, 2018; Simonsohn et al., 2020). Notably, most effects in the present study were robust to differences in data pre-processing: Either all 126 pathways indicated a significant result, or no pathway did, increasing confidence in the veracity of our findings.

We argue that the field of social cognition, and other research areas in which reaction time paradigms are frequently used, would heavily benefit from adopting specification curve analysis. For reaction time data, there often is no strong justification to adopt any single pre-processing pathway over others, and surveys of the published literature (Kerr et al., 2017; Primbs et al., 2022) and many analyst projects (Dutilh et al., 2019) indicate that even experts do not agree on what constitutes the right model, analysis or pre-processing pathway. Importantly, studies have shown that these pre-processing decisions heavily influence possible outcomes of statistical tests (André, 2021; Ratcliff, 1993) and determine the interpretation of any given study. Applying specification curve analysis provides researchers with more confidence in their findings, which strongly benefits scientific progress.

4.1. Limitations and future directions

Finally, our studies have at least two notable limitations. We investigated the effect of face masks on the stereotype effect using a speeded categorization paradigm only. While that is consistent with previous work demonstrating the stereotype effect (Bijlstra et al., 2010), it could be argued that face masks affect perception differently in dynamic emotion displays (Bijlstra et al., 2014). As such, the evidence provided in the present study is limited to specific types of emotion perception processes. Future studies should therefore expand the range of paradigms used to investigate stereotype effects in emotion perception and their underlying processes. In addition, the stimuli used in the present set of studies were created by adding face masks to existing images; and not by taking images of actors wearing face masks, who might display emotions differently than without face masks. This may have further complicated the perception of emotions on masked faces. However, different manipulations to decrease the strength of the visual input also did not produce the expected result (see Supplementary Materials S1). Still, future research should employ other, more potent manipulations.

To summarise, the present research investigated the effects of face masks on the stereotype effect in emotion perception. We successfully replicated male-anger and ethnicity-emotion stereotypes in emotion perception and demonstrated the efficacy of specification curve analysis in social cognition research but did not find evidence that face masks increase reliance on prior stereotype associations. Our findings challenge the applicability of predictive coding frameworks to social cognition research.

Open practices

Our studies were preregistered. The preregistration for Study 1 can be found here: https://osf.io/pz4xh/?view_only=c05a128203944b9e81fdf7e285ef909e. The preregistration for Study 2 is vastly identical to the one for Study 1 and can be found here: https://osf.io/6ayd5?view_only=bce1ec71085d40aeac7963ff7cf77b73. The data and analysis scripts associated with Study 1 and Study 2 and the supplementary materials can be found here: https://osf.io/vcwbu/files?view_only=eb2f0ee86614481eb7f354168963a40e.

Author contributions

M.P., M.R., & G.B. designed and executed Study 1 and Study 2. M.P. analysed the data and wrote the first draft of the manuscript. W.K., A.N., R.H., & G.B. designed and executed supplementary studies S1 and S2. W. K. and A.N. analysed the data of the supplementary studies and wrote the first draft of the supplementary materials. All authors commented on and approved the final draft.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

There are no conflicts of interest.

Appendix A. Software use statement

The present study relied on R (R Core Team, 2020) and the following software packages programmed in R: MASS (version 7.3.51.6; Venables

& Ripley, 2002), tidyverse (version 1.3.0; Wickham et al., 2019), plyr (version 1.8.6; Wickham, 2011), ggplot2 (version 3.3.3; Wickham, 2016), dplyr (version 1.0.2; Wickham et al., 2020, ggpubr (version 0.4.0; Kassambara, 2020), rstatix (version 0.6.0; Kassambara, 2020), here (version 0.1; Müller, 2017), reshape2 (version 1.4.4; Wickham, 2007), gtools (version 3.8.2; Warnes et al., 2020), ggthemes (version 4.2.0; Arnold, 2019), effsize (version 0.8.0; Torchiano, 2020), desctools (version 0.99.38; Carbon, 2020), rqPen (version 2.2.2; Sherwood & Maidman, 2020), data.table (version 1.13.0; Dowle & Srinivasan, 2020), ez (version 4.4.0; Lawrence, 2016), janitor (version 2.01.1; Firke, 2020), afex (version 0.28.1; Singmann et al., 2018), psycho (version 0.6.1; Makowski, 2018), and gridExtra (version 2.3; Auguie, 2017).

References

Arnold, J. B. (2019). ggthemes: extra themes, scales and geoms for 'ggplot2'. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes

Auguie, B. (2017). gridExtra: miscellaneous functions for "grid" graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Dowle, M., & Srinivasan, A. (2020). data.table: extension of `data.frame`. R package version 1.13.0. https://CRAN.R-project.org/package=data.table

Firke, S. (2020). janitor: simple tools for examining and cleaning dirty data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

Kassambara, A. (2020). ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Kassambara, A. (2020). rstatix: Pipe-friendly framework for basic statistical tests. R package version 0.6.0. https://CRAN.R-project.org/package=rstatix

Lawrence, M. A. (2016). ez: easy analysis and visualization of factorial experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez

Makowski, D. (2018). The psycho package: an efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software, 3*(22), 470. Available from https://github.com/neuropsychology/psycho.R

Müller, K. (2017). here: a simpler way to find your files. R package version 0.1. https://CRAN.R-project.org/package=here

R Core Team (2020). R. A language and environment for statistical computing. R Foundation for Statistical Computing.

Sherwood, B., & Maidman A. (2020). rqPen: penalized quantile regression. R package version 2.2.2. https://CRAN.R-project.org/package=rqPen

Signorell, A. (2020). DescTools: Tools for descriptive statistics. R package version 0.99.38. https://cran.r-project.org/web/packages/DescTools/index.html

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *afex: Analysis of factorial experiments*. R package version 0.28-1. Retrieved from https://CRAN.R-project.org/package=afex

Torchiano, M. (2020). _effsize: efficient effect size computation. R package version 0.8.0. https://CRAN.R-project.org/package=effsize

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S* (4th ed.). Springer.

Warnes, G. R., Bolker, B., & Lumley, T. (2020). gtools: various R programming tools. R package version 3.8.2. https://CRAN.R-project.org/package=gtools

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1-20. http://www.jstatsoft.org/v21/i12/.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*(1), 1-29. http://www.jstatsoft.org/v40/i01/.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L. Francois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Milton Bache, S., Müller, K., Ooms, J., Robinson, D., Paige Seidel, D., Spinu, V., Takashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., Francois, R., Henry, L., & Müller, K. (2020). dplyr: a grammar of data manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104394.

References

Amodio, D. M., & Devine, P. G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology, 91*(4), 652–661. https://doi.org/10.1037/0022-3514.91.4.652

André, Q. (2021). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001069. Advance online publication.

Ashwin, C., Chapman, E., Colle, L., & Baron-Cohen, S. (2006). Impaired recognition of negative basic emotions in autism: A test of the amygdala theory. *Social Neuroscience, 1*(3–4), 349–363. https://doi.org/10.1080/17470910601040772

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., & Wheelwright, S. (2003). The systemizing quotient: An investigation of adults with Asperger syndrome or high–functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1430), 361–374. https://doi.org/10.1098/rstb.2002.1206

Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K. C., & Smith, D. M. (2007). The confounded nature of angry men and happy women. *Journal of Personality and Social Psychology, 92*(2), 179–190. https://doi.org/10.1037/0022-3514.92.2.179

Bijlstra, G., Holland, R. W., Dotsch, R., Hugenberg, K., & Wigboldus, D. H. (2014). Stereotype associations and emotion recognition. *Personality and Social Psychology Bulletin, 40*(5), 567–577. https://doi.org/10.1177/0146167213520458

Bijlstra, G., Holland, R. W., Dotsch, R., & Wigboldus, D. H. J. (2019). Stereotypes and prejudice affect the recognition of emotional body postures. *Emotion, 19*(2), 189–199. https://doi.org/10.1037/emo0000438

Bijlstra, G., Holland, R. W., & Wigboldus, D. H. (2010). The social face of emotion recognition: Evaluations versus stereotypes. *Journal of Experimental Social Psychology, 46*(4), 657–663. https://doi.org/10.1016/j.jesp.2010.03.006

Blais, C., Roy, C., Fiset, D., Arguin, M., & Gosselin, F. (2012). The eyes are not the window to basic emotions. *Neuropsychologia, 50*(12), 2830–2838. https://doi.org/10.1016/j.neuropsychologia.2012.08.010

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods, 44*, 1255–1265.

Carbon, C. (2020). Wearing face masks strongly confuses counterparts in reading emotions. *Frontiers in Psychology, 11*, 66886. https://doi.org/10.3389/fpsyg.2020.566886

Chen, J. M., & Hamilton, D. L. (2012). Natural ambiguities: Racial categorization of multiracial individuals. *Journal of Experimental Social Psychology, 48*(1), 152–164. https://doi.org/10.1016/j.jesp.2011.10.005

Christensen, B., & Christensen, S. (2014). Are female hurricanes really deadlier than male hurricanes? *PNAS, 111*(344), E3497–E3498. https://doi.org/10.1073/pnas.1410910111

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass, 8*(5), 201–213. https://doi.org/10.1111/spc3.12099

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314

Craig, B. M., Koch, S., & Lipp, O. V. (2017). The influence of social category cues on the happy categorisation advantage depends on expression valence. *Cognition and Emotion, 31*(7), 1493–1501. https://doi.org/10.1080/02699931.2016.1215293

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science, 4*(1), 1–15. https://doi.org/10.1177/2515245920954925

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Dotsch, R., & Wigboldus, D. H. (2008). Virtual prejudice. *Journal of Experimental Social Psychology, 44*(4), 1194–1198. https://doi.org/10.1016/j.jesp.2008.03.003

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., … Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review, 26*(4), 1051–1069. https://doi.org/10.3758/s13423-017-1417-2

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology, 87*(6), 876–893. https://doi.org/10.1037/0022-3514.87.6.876

Feldman Barrett, L., Adolphs, R., Marsella, S., Martinzed, 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. https://doi.org/10.1177/1529100619832930

Gayet, S., & Stein, T. (2017). Between-subject variability in the breaking continuous flash suppression paradigm: Potential causes, consequences, and solutions. *Frontiers in Psychology, 8*, 437. https://doi.org/10.3389/fpsyg.2017.00437

Harms, M. B., Martin, A., & Wallace, G. L. (2010). Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychology Review, 20*(3), 290–322. https://doi.org/10.1007/s11065-010-9138-6

Hugenberg, K. (2005). Social categorization and the perception of facial affect: Target race moderates the response latency advantage for happy faces. *Emotion, 5*(3), 267–276. https://doi.org/10.1037/1528-3542.5.3.267

Hugenberg, K., & Sczesny, S. (2006). On wonderful women and seeing smiles: Social categorization moderates the happy face response latency advantage. *Social Cognition, 24*(5), 516–539. https://doi.org/10.1521/soco.2006.24.5.516

Hutchings, P. B., & Haddock, G. (2008). Look black in anger: The role of implicit prejudice in the categorization and perceived emotional intensity of racially ambiguous faces. *Journal of Experimental Social Psychology, 44*(5), 1418–1420. https://doi.org/10.1016/j.jesp.2008.05.002

Inquisit 5 [Computer Software]. Retrieved from https://www.millisecond.com, (2016).

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014a). Female hurricanes are deadlier than male hurricanes. *PNAS, 111*(24), 8782–8787. https://doi.org/10.1073/pnas.1402786111

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014b). Reply to Christensen and Christensen and to Malter: Pitfalls of erroneous analyses of hurricanes names. *PNAS, 111*(34), E3499–E3500. https://doi.org/10.1073/pnas.1402786111

Kerr, J. A., Hesselmann, G., Räling, R., Wartenburger, I., & Sterzer, P. (2017). Choice of analysis pathway dramatically affects statistical outcomes in breaking continuous flash suppression. *Scientific Reports, 7*, 3002. https://doi.org/10.1038/s41598-017-03396-3

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., & Van Knippenberg, A. (2010). Presentation and validation of the Radboud faces database. *Cognition & Emotion, 24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Malter, D. (2014). Female hurricanes are not deadlier than male hurricanes. *PNAS, 111*(34), E3496. https://doi.org/10.1073/pnas.1411428111

Marsh, A. A., Adams, R. B., Jr., & Kleck, R. E. (2005). Why do fear and anger look the way they do? Form and social function in facial expressions. *Personality and Social Psychology Bulletin, 31*(1), 73–86. https://doi.org/10.1177/0146167204271306

Munoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology, 48*(1), 1–33. https://doi.org/10.1177/0081175018777988

Neth, H., & Gigerenzer, G. (2015). Heuristics: Tools for an uncertain world. In R. A. Scott, & S. M. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences*. Wiley. https://doi.org/10.1002/9781118900772.etrds0394.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*(3), 466–478. https://doi.org/10.1037/0096-3445.130.3.466

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80*(3), 381–396. https://doi.org/10.1037/0022-3514.80.3.381

Otten, M., Seth, A. K., & Pinto, Y. (2017). A social Bayesian brain: How social knowledge can shape visual perception. *Brain and Cognition, 112*, 69–77. https://doi.org/10.1016/j.bandc.2016.05.002

Pallier, C. (2002). Computing discriminability and bias with the R software. Retrieved from: https://www.pallier.org/pdfs/aprime.pdf.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248. https://doi.org/10.1080/1047840X.2017.1335568

Plant, E. A., Hyde, J. S., Keltner, D., & Devine, P. G. (2000). The gender stereotyping of emotions. *Psychology of Women Quarterly, 24*(1), 81–92. https://doi.org/10.1111/j.1471-6402.2000.tb01024.x

Primbs, M. A., Holland, R., Quandt, J., & Bijlstra, G. (2022). Data pre-processing distorts results and conclusions in reaction time data. Retrieved from https://osf.io/9gj8v/?view_only=9a998522319c4a38ba5dc302815a04fe.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Rinck, M., Primbs, M. A., Verpaalen, I., & Bijlstra, G. (2022). The effects of face masks on facial emotion recognition. Retrieved from https://osf.io/bjhct/.

Sacco, D. F., & Hugenberg, K. (2009). The look of fear and anger: Facial maturity modulates recognition of fearful and angry expressions. *Emotion, 9*(1), 39–49. https://doi.org/10.1037/a0014081

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science, 16*(3), 184–189. https://doi.org/10.1111/j.0956-7976.2005.00801.x

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Tamietto, M., & de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience, 11*, 697–709. https://doi.org/10.1038/nrn2889

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases: Biases in judgement reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Verkuyten, M., & Zaremba, K. (2005). Interethnic relations in a changing political context. *Social Psychology Quarterly, 68*(4), 375–386. https://doi.org/10.1177/019027250506800405