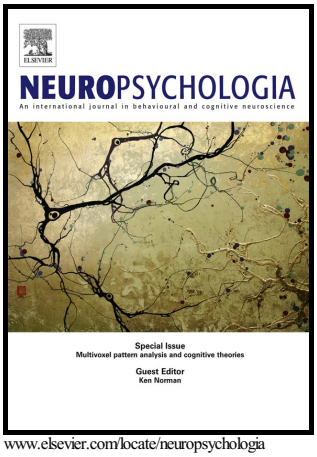
Author's Accepted Manuscript



www.elsevier.com/locate/neuropsychologia

Asymmetric neural responses for facial expressions and anti-expressions

O. Scott Gwinn, Courtney N. Matera, Sean F. O'Neil, Michael A. Webster

PII: S0028-3932(18)30554-2
DOI: https://doi.org/10.1016/j.neuropsychologia.2018.09.001
Reference: NSY6905

To appear in: *Neuropsychologia*

Received date: 9 June 2018
Revised date: 27 August 2018
Accepted date: 2 September 2018

Cite this article as: O. Scott Gwinn, Courtney N. Matera, Sean F. O'Neil and Michael A. Webster, Asymmetric neural responses for facial expressions and anti-expressions, *Neuropsychologia,* https://doi.org/10.1016/j.neuropsychologia.2018.09.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Asymmetric neural responses for facial expressions and anti-expressions.

O. Scott Gwinn[a,b,1], Courtney N. Matera[a,1], Sean F. O'Neil[a], and Michael A. Webster[a]

[a]Department of Psychology, University of Nevada, Reno, USA.

1664 N Virginia St, Reno, NV, 89557

[b]School of Psychology, Flinders University, Adelaide, Australia.

Sturt Rd, Bedford Park, South Australia 5042

[1]The first two authors contributed equally

Correspondences should be addressed to O. Scott Gwinn (dr.scottgwinn@gmail.com)

Accepted manuscript

ACCEPTED MANUSCRIPT

Abstract

Face recognition requires identifying both the invariant characteristics that distinguish one individual from another and the variations within the individual that correspond to emotional expressions. Both have been postulated to be represented via a norm-based code, in which identity or expression are represented as deviations from an average or neutral prototype. We used Fast Periodic Visual Stimulation (FPVS) with electroencephalography (EEG) to compare neural responses for neutral faces, expressions and anti-expressions. Anti-expressions are created by projecting an expression (e.g. a happy face) through the neutral face to form the opposite facial shape (anti-happy). Thus expressions and anti-expressions differ from the norm by the same "configural" amount and thus have equivalent but opposite status with regard to their shape, but differ in their ecological validity. We examined whether neural responses to these complementary stimulus pairs were equivalent or asymmetric, and also tested for norm-based coding by comparing whether stronger responses are elicited by expressions and anti-expressions than neutral faces. Observers viewed 20 s sequences of 6 Hz alternations of neutral faces and expressions, neutral faces and anti-expressions, and expressions and anti-expressions. Responses were analyzed in the frequency domain. Significant responses at half the frequency of the presentation rate (3 Hz), indicating asymmetries in responses, were observed for all conditions. Inversion of the images reduced the size of this signal, indicating these asymmetries are not solely due to differences in the low-level properties of the images. While our results do not preclude a norm-based code for expressions, similar to identity, this representation (as measured by the FPVS EEG responses) may also include components sensitive to which configural distortions form meaningful expressions.

Keywords: Facial expressions; Norm-based coding; EEG; FPVS

Accepted manuscript

ACCEPTED MANUSCRIPT

1. Introduction

Expressions are integral to humans' interactions with their environment, serving as markers of emotional responses and facilitating social communication (Eimer, Holmes, & McGlone, 2003; Fridlund, 1997; Russell & Fernández-Dols, 1997). Certain expressions are considered to be universal, in that the emotions associated with them are agreed upon cross-culturally (Ekman, 1992; Izard, 1992; for an alternative view see Russell, 1994). For the past four decades researchers have typically focused on six basic expressions: disgust, fear, happiness, sadness, surprise and anger. These expressions presumably evolved to serve different functional purposes, such as avoiding noxious stimuli in the case of disgust (Rozin, Haidt, & McCauley, 2008). Considerable research has been devoted to understanding the recognition of expressions and their neural underpinnings (Calvo & Nummenmaa, 2016). However, the basic coding strategies for representing expressions remain uncertain.

One proposed strategy for representing facial information is norm-based coding. In norm-based coding individual exemplars are represented according to how they deviate from a central prototype. This is analogous to an opponent color space, in which chromatic hue and saturation can be represented as vector directions and lengths from a central gray (Hurvich & Jameson, 1957; Webster & MacLeod, 2011). Norm-based coding has been widely (though not universally) postulated as the representational architecture for facial identity (Rhodes & Leopold, 2011; Valentine, Lewis, & Hills, 2016; Webster & MacLeod, 2011). This can be conceptualized as a hypothetical 'face space' in which the average of previously encountered faces forms the central norm and more distinctive faces are encoded as points further from the norm (Valentine, 1991). Faces can be made to appear more distinctive through caricaturing, which exaggerates the physical differences between an original face and an average norm. Anti-caricaturing involves reducing these differences, thus making a face appear more typical (Leopold, O'Toole, Vetter, & Blanz, 2001). Norm-based models predict relatively low levels of neurophysiological activity for typical faces and monotonically increasing levels of activity for more distinctive faces (Loffler, Yourganov, Wilkinson, & Wilson, 2005). Using functional magnetic resonance imaging (fMRI), Loffler et al. (2005) demonstrated that larger blood oxygenation level-dependent (BOLD) signals are associated with distinctive faces compared to average faces. Leopold, Bondar and Giese (2006) reported similar monotonic response increases with identity strength for single neurons in the monkey inferotemporal cortex. In electroencephalography (EEG) recordings, event-related potentials (ERP's) with peak latencies between 200 and 250 ms, known as the P200 component, appear to also reflect the typicality of faces (Kaufmann & Schweinberger, 2012; Schulz, Kaufmann, Kurt, & Schweinberger, 2012; Schulz, Kaufmann, Walther, & Schweinberger, 2012). In the case of the P200, larger amplitudes are observed for more typical faces compared to distinctive faces (Schulz et al., 2012), likely due to the ERP component reflecting different processes than the BOLD signal. Regardless, the modulation of the component by caricaturing is consistent with a special status of the norm.

Behavioral studies of face adaptation and the resulting aftereffects have also implicated norm-based coding for many aspects of faces (Jiang, Blanz, & O'Toole, 2006; Leopold et al., 2001; Rhodes & Jeffery, 2006). For example, adaptation is stronger for a distorted than an undistorted face, potentially because the former recalibrates the norm while an undistorted adaptor instead reinforces the norm (Webster & MacLin, 1999). A role for norms has also been implicated by adaptation to "anti-faces," in which the physical characteristics of an original identity are warped through a mean/norm and 'out the other side,'

Accepted manuscript

ACCEPTED MANUSCRIPT

resulting in face that is opposite in configuration to the original, analogous to complementary colors in color space (Blanz, O'Toole, Vetter, & Wild, 2000). Adapting to an anti-face results in a norm appearing more similar to the original identity (Jiang et al., 2006; Leopold et al., 2001; Rhodes & Jeffery, 2006), and such effects have been interpreted in terms of two pools of neural mechanisms whose relative responses encode the intensity or distinctiveness of the face along a given identity trajectory (Rhodes et al., 2005; Webster & MacLeod, 2011). Recent studies have also demonstrated face adaptation effects in EEG responses (Retter & Rossion, 2016b) and have shown that these implicate a renormalization of face processing (Kloth, Rhodes, & Schweinberger, 2017).

Analogous norm-based accounts have also been proposed for the representation of facial expressions. Conceptually, some emotions are portrayed as "opposites," and expressions have been ordered as different directions within a space defined by two general evaluative dimensions where the center is a neutral expression (e.g. Russell, 1980). Perceptually, a norm-based code has also been postulated based again on adaptation aftereffects, which like identity aftereffects can lead to large biases in perceived expression (Benton, 2009; Benton et al., 2007; Fox & Barton, 2007; Hsu & Young, 2004; Rutherford, Chattha, & Krysko, 2008; Webster, Kaping, Mizokami, & Duhamel, 2004). For example, when two expressions are morphed to create an intermediate image, adapting to one of the expressions causes the intermediate image to appear more like the other expression (Webster et al., 2004). Also like face identity, expression aftereffects have been probed by adapting to anti-expressions (Cook, Matei, & Johnston, 2011; Juricevic & Webster, 2012; Rhodes et al., 2017; Skinner & Benton, 2010, 2012). These anti-expressions are created by warping an expression through a mean (either a neutral expression or the average of many expressions) to create a configural opposite (see Figure 1). Adapting to an anti-expression biases subsequent perceptions towards the original expression, and this has been taken as evidence that expressions and anti-expressions form opponent pairs.

However, there are fundamental differences between the physical variations defining identity and expression which potentially limit the application of norm-based coding to expressions. Within an identity space, vectors radiating along many directions from a central norm are likely to produce biologically plausible identities. Conversely, in an expression space only a finite number of vectors will correspond to plausible muscle distortions, and a smaller number still will correspond to meaningful expressions. This poses a problem because the norm is presumably set by the average face perceived, yet expressions and their anti-expressions should be expected to occur with very different frequencies. Moreover, in the case of expressions it is not clear what the norm or prototype should be. In some cases this has been interpreted to reflect the average of all expressions (e.g. Cook et al., 2011; Skinner and Benton, 2010, 2012). In that case there may be plausible faces on all sides of the mean, yet the mean itself is not necessarily the facial configuration that corresponds to a "neutral" expression (Juricevic & Webster, 2012).

Here we used neurophysiological responses associated with processing facial expressions to examine the extent to which expressions are encoded in a norm-based fashion. Studies of the neural responses to expressions have typically used ERP's to examine the time course of encoding (Blau, Maurer, Tottenham, & McCandliss, 2007; Caharel, Courtay, Bernard, Lalonde, & Rebaï, 2005; Eimer & Holmes, 2002; Schupp et al., 2006; Stefanics, Csukly, Komlósi, Czobor, & Czigler, 2012). While there is some inconsistency in results across studies, those finding stronger responses for basic expressions than

ACCEPTED MANUSCRIPT

neutral faces (Luo, Feng, He, Wang, & Luo, 2010; Williams, Palmer, Liddell, Song, & Gordon, 2006) are consistent with norm-based coding tests of face identity. However, it is not known how these neural responses might differ for expressions and their anti-expressions, and thus whether such results can be explained in terms of a representation involving an actual norm.

The differences in ERP waveforms to different expressions are often subtle (Calvo & Nummenmaa, 2016; Dzhelyova, Jacques, & Rossion, 2016). However, researchers have recently begun examining the neural correlates of face processing using an alternative paradigm known as Fast Periodic Visual Stimulation (FPVS), in which images are presented at a fixed rate and responses analyzed in the frequency domain (Rossion, 2014a, 2014b). This technique is based on earlier observations that the human brain will synchronize its activity with a flickering stimulus, a process more commonly known as Steady-State Visual Evoked Potentials (SSVEP) (Regan, 1966; for review see Norcia, Appelbaum, Ales, Cottereau, & Rossion, 2015). This allows responses to the stimulus of interest to be tracked by "frequency tagging" or isolating the response amplitude and phase at the relevant temporal frequency. FPVS has many advantages, such as its objectivity (i.e. responses occur at an experimenter defined frequency), its implicit nature (i.e. it does not require any additional tasks), its resistance to artifacts, which allows for the production of high signal-to-noise ratios with few trials, and the relative simplicity of data analysis and interpretation (Rossion, 2014a, 2014b). The effectiveness of FPVS for studying expressions was recently demonstrated by Dzhelyova et al. (2016). Using an 'oddball' paradigm, in which an expression was periodically presented within a sequence of neutral faces, Dzhelyova et al. (2016) showed that the brain could rapidly discriminate between neutral and expressive faces, and that these responses were driven by high-level aspects of the faces and not simply low-level differences in the images.

In the present study we used the FPVS paradigm to explore the relative responses to expressions, anti-expressions, and neutral faces, and the implications of these responses for the representation of facial expressions. Specifically, we asked whether expressions and anti-expressions shared an equivalent status relative to a neutral face. If expressions and anti-expressions are equivalently represented as deviations from a neutral norm (at the level driving the FPVS response) then we should expect the two stimuli to produce equal responses, and that both expressions and anti-expressions should yield larger responses than neutral faces. To test these predictions we applied the FPVS technique in a paradigm modelled after Retter and Rossion (2016), in which two facial images are sequentially presented at a fixed rate (6 Hz), and asymmetries in the response generated by these images are indicated by the presence of a signal at the half harmonic (3 Hz). Whereas Retter and Rossion (2016) focused on asymmetries in responses to different identities following adaptation to one of the identities, we tested for the presence of inherent asymmetries in the processing of neutral faces, expressions and anti-expressions. We also evaluated these effects for two classes of images: 1) facial photographs of individuals posing in different expressions (to compare responses to actual expressive vs. neutral faces); and 2) computed generated images, which allowed us to present both expressions and the morphed anti-expressions.

2. Methods

ACCEPTED MANUSCRIPT

2.1 Participants

Two of the authors (OG and CM) and twelve undergraduate psychology students at the University of Nevada, Reno (14 total, 8 male) with a mean age of 23.4 (*SD* = 3.4) took part in the study. Optical corrections were worn if required. Participation was with informed consent and followed protocols approved by the university's Institutional Review Board.

2.2 Stimuli

As noted, we used two sets of face images, each with different advantages. The face photographs allowed us to measure responses to photo-realistic images of expressions to ensure the ecological validity of the measurements. The computer generated images appear less realistic but have the advantage of providing parametric control over the facial identity and configuration, and the ability to generate different magnitudes of expressions and their anti-expressions. Photo-realistic images of faces were obtained from the Radboud Face Database (Langner, Dotsch, Bijlstra, Wigboldus, Hawk, & van Knippenberg, 2010). One Caucasian male and one Caucasian female model were chosen showing the expressions happy, anger, and fear, as well as a neutral expression (see Figure 1). Images were cropped using a standard ellipse to remove information outside of the face. Multiple expressions and male and female images were used to help ensure that any observed effects could not simply be attributed to the properties of a single image and instead reflect more general processing of expressions. Computer generated facial images were created using the FaceGen Modeller 3.3. This program has been widely used to create stimuli for experimental studies of facial expression perception (Clément, Bernard, Grandjean, & Sander, 2013; Engell, Todorov, & Haxby, 2010; Juricevic & Webster, 2012; N'diaye, Sander, & Vuilleumier, 2009). An average Caucasian male and female face were generated with neutral expressions. These head models were then varied using sliding scales available in the program, to display the expressions happy, anger and fear. The computer generated expressions thus differed from the photographs in a number of regards, chosen in each case to optimize the images within the constraints of the stimulus set. For example, while the "angry" photographs had closed mouths, for the simulated faces we used an open mouth since the anger expression was much more evident in these images. While open-mouth expressions can be perceived as more intense than closed-mouth (Horstmann, Lipp, & Becker, 2012), both forms are regarded as accurate representations of the emotion (Alvarado & Jameson, 2002). Corresponding anti-expression images were also created by moving the scales to the opposite side of the neutral starting point (see Figure 1). Images were cropped along the jaw and outer line of the head to remove information outside of the face. All images (22 total) were formatted to a height of approximately 300 pixels (approximately because overall height varies slightly due to jaw movements associated with different expressions) and placed on a standard grey background. These were presented on a NEC AccuSync 120 monitor with a working resolution of 1280 x 960 pixels and a refresh rate of 60 Hz. At a viewing distance of 57 cm the images subtended a visual angle of approximately 9.6 degrees.

Accepted manuscript

ACCEPTED MANUSCRIPT

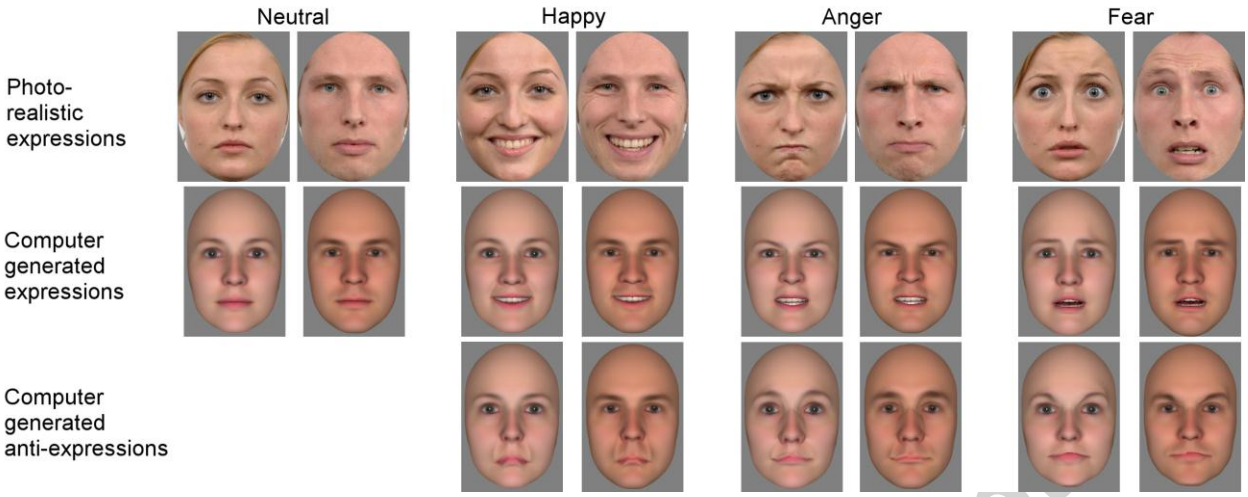


Figure 1. Facial images used in the study. Top row shows photo-realistic male and female images displaying the expressions neutral, happy, anger, fear. Middle row shows computer generated images displaying the same expressions. Bottom row shows the corresponding anti-expressions.

2.3 Procedure

The experiment took place in a quiet, darkened room. Each session lasted for approximately 1.5 hrs, including 30 min of preparation and 1 hr of recording. Images were presented using the FPVS technique (Rossion, 2014a, 2014b) and custom software running over Java 8 (Oracle, USA). A single trial consisted of a 20 s sequence in which stimuli were shown at a fixed rate of six images per second (6 Hz) by means of a square wave modulation with a gray background at a 50% duty cycle. Each sequence comprised two images alternating with each other, resulting in repetition of the same face at a frequency of 3 Hz. To reduce the potential impact of low-level properties of the images on the responses, the size of the images varied randomly across five steps between 90% and 110% of the original image size at each stimulus presentation cycle (Dzhelyova & Rossion, 2014). During each trial, participants were required to fixate on a cross in the center of the screen. To help ensure attention was maintained during each trial, the fixation cross would briefly change to a square eight times during each sequence at random intervals and participants were required to press a key to indicate when they saw the change.

The experiment was divided into six conditions defined by the two images that alternated. The first condition comprised alternations between photo-realistic images of a neutral face and an expression. In the second condition these images were presented inverted. The third condition comprised alternations between computer generated images of a neutral face and an expression, the fourth between a neutral face and an anti-expression, the fifth an expression and its anti-expression. The sixth condition was a repetition of the fifth but with the images inverted. Each condition contained 12 trials and within a trial only one gender and expression type were seen (two genders * three expressions * two repetitions = 12 trials). The experiment was divided into two blocks and within each block trials from each condition were randomly intermixed. As the neutral face was necessarily seen in more trials than each expression and anti-expression, participants may have become adapted to these images, potentially inducing

ACCEPTED MANUSCRIPT

asymmetries in responses (Retter & Rossion, 2016b). To control for this an additional 48 trials of the fifth condition (expression/anti-expression) as well as 6 trials in which two photo-realistic expressions alternated were randomly intermixed with the other trials, resulting in a total of 132 trials for the whole experiment. Data from the additional control trials were not analyzed.

2.4 EEG acquisition

The data were recorded using a BioSemi ActiveTwo system with a 128 Ag-AgCl Active-electrode array (BioSemi B.V., Amsterdam, Netherlands; for exact position coordinates, see http://www.biosemi.com/headcap.htm, for a conversion of these coordinates to a more standard 10-5 nomenclature (Oostenveld & Praamstra, 2001), see Rossion, Torfs, Jacques, & Liu-Shuang, 2015). Electrode offsets were kept below 40mV, referenced to the common mode sense (CMS). Four additional electrodes were used to record vertical and horizontal electrooculogram (EOG): two electrodes were placed above and below participants' right eye and two were placed lateral to the external canthi. The EEG and EOG were digitized at a sampling rate of 2048 Hz and then down-sampled to a rate of 512 Hz.

2.5 Analysis

The recorded EEG was analyzed using Letswave 5, an open source toolbox (http://nocions.webnode.com/letswave), running over MATLAB R2013b (MathWorks, USA).

2.5.1 Preprocessing

Data files for each participant were first filtered using a fourth order zero-phase Butterworth band-pass filter, with cutoff values of .1-120 Hz. A Fast Fourier Transform (FFT) multi-notch filter with a width of .5 Hz was also applied to remove electrical noise at three harmonics of 60 Hz. The data were then segmented by trial, including 1 s before and after the beginning of stimulation. To correct for artifacts caused by eye blinks, independent component analysis (ICA) with a square matrix was applied (Hyvarinen & Oja, 2000). A single component was removed for three participants who blinked more than .2 times/s on average during the 20 s stimulation sequences. The cutoff of .2 times/s was chosen based on Retter and Rossion (2016). Channels containing artifacts across multiple trials were replaced with the average of 3-4 neighboring channels. This was performed for a maximum of three channels per participant. All channels were then re-referenced to the common average. For each subject, trials were re-segmented to exclude the 1 s before and after the 20 s stimulation sequence. Trials were then averaged within each condition.

2.5.2 Frequency analysis

An FFT was computed for each subject, condition and channel. These data were then grand averaged across all subjects. Recordings were primarily analyzed using a right occipito-temporal (ROT) region of

Accepted manuscript

ACCEPTED MANUSCRIPT

interest (ROI), comprising electrodes PO8, PO10, PO12, P10 and P8. This region has previously been shown to be sensitive to neurophysiological responses associated with face processing (Dzhelyova & Rossion, 2014; Retter & Rossion, 2016). Some comparisons are also performed between the signal at this ROI and a medial occipital (MO) region, comprising electrodes PPOz, POz, POOz, Oz, and Oiz. The MO ROI has been shown to be less sensitive to neurophysiological responses associated with face processing and instead may reflect process associated more with the early stages of visual processing (Dzhelyova & Rossion, 2014; Retter & Rossion, 2016).

The presence of a significant response at the frequency of interest was determined by z-scores (z = (x-baseline)/standard deviation of the baseline). Baselines were defined as the twenty bins surrounding the bin of interest (x), excluding the immediately adjacent bins (Rossion, Prieto, Boremanse, Kuefner, & Belle, 2012; Srinivasan, Russell, Edelman, & Tononi, 1999). When displaying the amplitude spectra and comparing differences in amplitude across conditions, baseline corrections were applied to account for differences in baseline noise across participants and across the frequency spectrum within participants. This took the form of a baseline subtraction in which the average of the twenty surrounding bins, excluding the immediately adjacent bins and the local maximum and minimum amplitude bins, was subtracted from the bin of interest (x'= x-baseline). When comparing differences in amplitude, the sum of baseline-subtracted harmonics of the frequency of interest was also computed (see Heinrich, 2009). For responses at 3 Hz the even harmonics were not included as these correspond with the presentation rate of 6 Hz. Following Dzhelyova et al. (2016), who summed harmonics up to a frequency of 16.46 Hz for the 3 Hz signal and up to 17.64 Hz for the 6 Hz signal, we summed harmonics up to 15 Hz and 18 Hz respectively.

3. Results

In the first three sections of the results we examine the amplitude of the signal at 3 Hz across conditions. Being half the frequency of the presentation rate, significant responses at this frequency indicate an asymmetry in responses. In the fourth section we consider the phase of these responses. In the fifth section we provide a brief examination of responses at the presentation rate of 6 Hz. All analyses refer to responses recorded in the ROT ROI unless otherwise stated. Continuously significant odd harmonics with $p < .05$ were observed up to the 5th harmonic (15 Hz) in three conditions in the ROT ROI and four conditions in the MO ROI. Baseline-subtracted amplitudes for these harmonics were summed for each condition and region separately before making comparisons between conditions and regions.

As the aim of the present study is to examine a global account of expression processing and not potential differences in the processing of different facial genders and expressions, data for male and female images and the three expressions are combined in all proceeding analyses. However, prior to these analyses we subjected the data to a 2 x 3 repeated measures ANOVA with the factors 'gender' (male vs. female) and 'expression' (happy, fear, anger) to confirm that the amplitude of responses at 3 Hz did not significantly differ for faces of different gender ($F_{1,13} = .08$, $p = .78$, $\eta_p^2 = .01$) or displaying different expressions ($F_{2,26} = 1.91$, $p = .169$, $\eta_p^2 = .13$). The gender*expression interaction was also non-significant ($F_{2,26} = .11$, $p = .898$, $\eta_p^2 = .01$).

ACCEPTED MANUSCRIPT

3.1 Expressions and neutral faces

We first sought to confirm for our conditions the results previously reported by Dzhelyova et al. (2016) showing that the FPVS paradigm can be used to observe asymmetries in the neural responses for photo-realistic images of expressions and neutral faces. We also sought to confirm that these results could be replicated using computer generated images. For the photo-realistic images, significant responses at 3 Hz were observed (z = 12.12, p <.001, one-tailed), confirming that the expressions and neutral faces produced asymmetrical responses (see Figure 2). For the computer generated images, significant responses at 3 Hz were also observed (z = 6.16, p <.001, one-tailed), indicating that these images were also sufficient in producing asymmetric responses (see Figure 2). Scalp topographies for both these conditions can be seen in Figure 3. For both image types, significant responses at 3 Hz were also evident for the majority of participants at the individual level (see Table 1). A comparison of the summed harmonics of the responses produced by the two image types showed that responses for photo-realistic images (M = .44, SD = .24) were significantly larger than responses for computer generated images (M = .2, SD = .11) (t_{13} = 3.26, p = .006, d = 1.24). This suggests that while computer generated images are sufficient to produce asymmetries in responses to expressions and neutral faces, the more naturalistic appearance of the photo-realistic images may better activate face encoding mechanisms and produce larger asymmetries in the responses in these areas. Nevertheless, that asymmetries in responses for expressions and neutral faces were observed using computer generated images indicates they are appropriate to use for addressing the aims of the current study.

Accepted manuscript

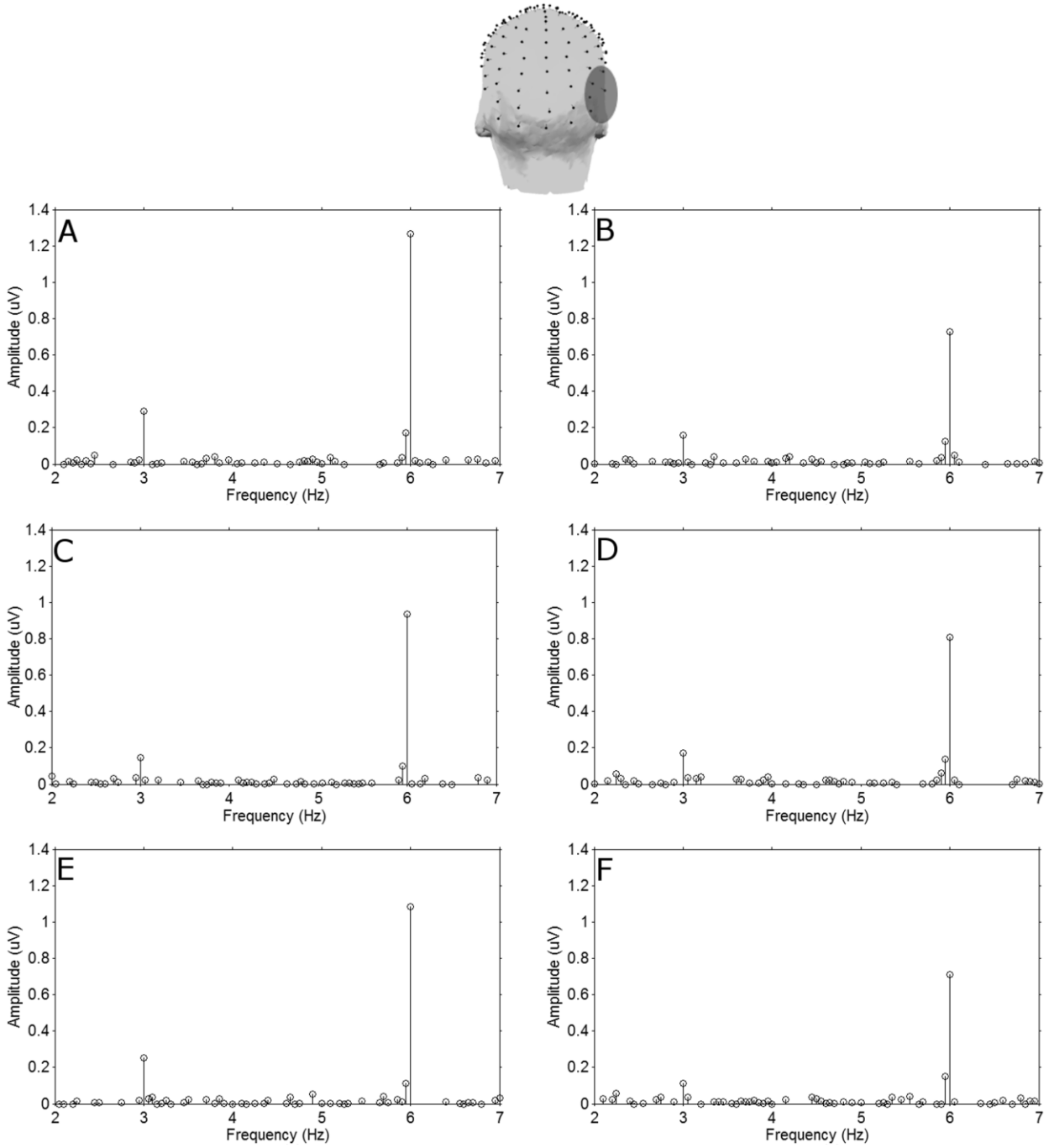


Figure 2. Top scalp map shows the electrodes comprising the ROT ROI. (A) Amplitude spectra for the sequence of alternating neutral/expression photo-realistic images, (B) spectra when these images were inverted, (C) spectra for the sequence of neutral/expression computer generated images, (D) the neutral/anti-expression sequence, (E) the expression/anti-expression sequence and (F) spectra when the expression/anti-expression images were inverted.

ACCEPTED MANUSCRIPT

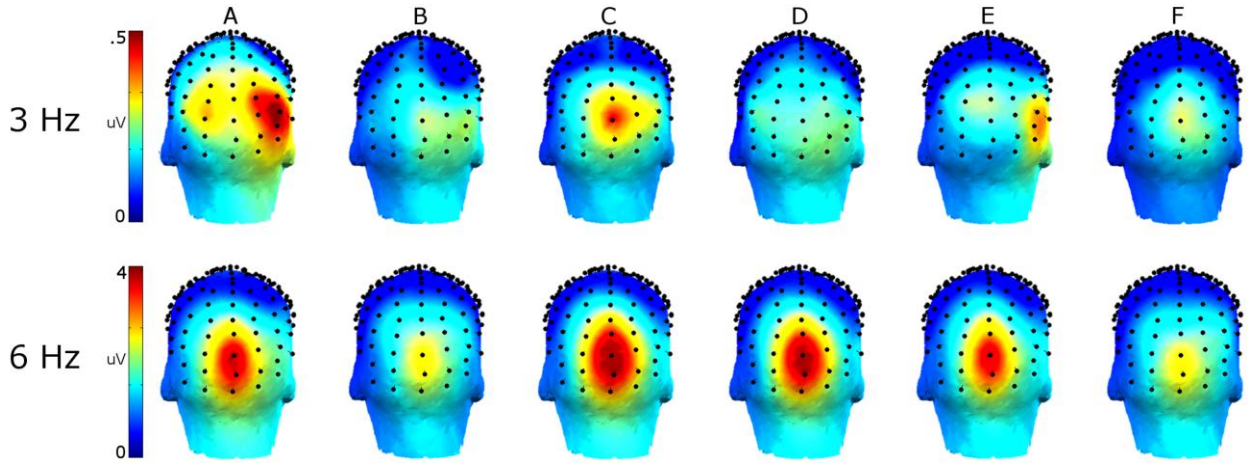


Figure 3. Scalp topographies summed across the first five odd harmonics for responses at 3 Hz (top) and summed across the first three harmonics for 6 Hz (bottom). Column (A) topographies for the sequence of alternating neutral/expression photo-realistic images, (B) topographies when these images were inverted, (C) topographies for the sequence of neutral/expression computer generated images, (D) the neutral/anti-expression sequence, (E) the expression/anti-expression sequence and (F) topographies when the expression/anti-expression images were inverted.

Table 1. Participants' individual z-scores for responses at 3 Hz in the ROT ROI. Rows show data for each participant and columns show data for each condition, with (A) corresponding to the sequence of alternating neutral/expression photo-realistic images, (B) when these images were inverted, (C) the sequence of neutral/expression computer generated images, (D) the neutral/anti-expression sequence, (E) the expression/anti-expression sequence and (F) when the expression/anti-expression images were inverted.

A	B	C	D	E	F
2.16 *	6.88 ***	5.74 ***	3.04 ***	6.65 ***	3.49 ***
8.47 ***	4.92 ***	1.30	1.20	6.74 ***	3.98 ***
2.69 **	2.17 *	-0.28	6.35 ***	4.14 ***	-0.29
1.75 *	2.11 *	0.66	-0.34	3.47 ***	0.69
3.01 **	0.40	3.65 ***	1.39	1.06	0.80
8.47 ***	4.91 ***	1.30	1.19	6.75 ***	3.97 ***
8.08 ***	3.14 ***	6.02 ***	1.06	3.21 ***	0.41
4.04 ***	4.06 ***	0.97	1.24	0.92	2.13 *
5.49 ***	0.95	2.38 **	1.17	0.85	2.14 *
9.31 ***	2.83 **	1.13	2.47 **	3 **	-0.98
4.42 ***	1.37	3.71 ***	9.88 ***	17.41 ***	4.40 ***
-0.01	0.28	4.70 ***	4.27 ***	1.19	3.22 ***
6.88 ***	4.65 ***	5.91 ***	6.96 ***	1.23	-1.00
8.57 ***	1.29	2.42 **	-0.13	2.08 *	2.97 **

* $p < .05$, ** $p < .01$, *** $p < .001$ (one tailed)

3.2 Anti-expressions

Having shown asymmetries in the responses for computer generated expressions and neutral faces, we next wished to examine how responses to these images compare to responses for anti-expressions. For

ACCEPTED MANUSCRIPT

alternations between anti-expressions and neutral faces a significant signal at 3 Hz was observed (z = 6.86, p <.001, one-tailed), indicating asymmetries in the responses to these images, similar to the previously reported results for expressions and neutral faces (see Figure 2). For alternations between anti-expressions and expressions a significant signal at 3 Hz was also observed (z = 15.56, p <.001, one-tailed), indicating that while these images are equal in terms of the extent to which they differ configurally from a neutral face, they produce asymmetric responses (see Figure 2). Scalp topographies for both these conditions can be seen in Figure 3.

3.3 Face specificity of asymmetries

To examine whether the asymmetries reported above reflect differences in face processing vs. low-level feature properties, we compared the responses for the same stimuli when the faces were inverted. Inversion of facial images has been shown to disrupt normal face processing in both behavioral (Farah & Tanaka, 1995; Freire, Lee, & Symons, 2000) and neuroimaging studies (Kanwisher, Tong, & Nakayama, 1998; Rossion, Delvenne, Debatisse, & Goffaux, 1999). Dzhelyova et al. (2016) previously demonstrated that inversion of photo-realistic images of expressions and neutral faces reduces asymmetries in the responses for these images, supporting the conclusion that these effects are partly due to higher-level processes associated with face encoding. These effects have also been shown to be more prominent over occipito-temporal sites compared to more medial occipital sites, thus validating the former as an appropriate site for examining responses related to face processing (Dzhelyova et al., 2016). We first sought to confirm the results reported by Dzhelyova et al. (2016) for our stimuli, by inverting our photo-realistic images of expressions and neutral faces and comparing responses across the ROT ROI and MO ROI. Significant responses at 3 Hz were observed in both the ROT ROI (z = 8.31, p <.001, one-tailed) and MO ROI (z = 10.36, p <.001, one-tailed) for inverted images (see Figure 2 and Figure 4). However, for the ROT ROI the difference between response amplitudes for upright (M = .43, SD = .24) and inverted (M = .25, SD = .17) images was significant (t_{13} = 3.97, p = .002, d = .91), while in the MO ROI the difference between upright (M = .27, SD = .23) and inverted (M = .21, SD = .13) images was not significant (t_{13} = .88, p = .395, d = .33). These results indicate the asymmetries in responses for photo-realistic images of expressions and neutral faces reported in the current study can partly be attributed to face related processes beyond the early stages of encoding. Further, that this difference was only evident in the ROT ROI supports the selection of this region for examining processes related to face encoding.

We next sought to apply these same tests to the asymmetries observed in the responses for computer generated expressions and anti-expressions by inverting these images. Significant responses at 3 Hz were observed in both the ROT ROI (z = 5.67, p <.001, one-tailed) and MO ROI (z = 9.45, p <.001, one-tailed) for inverted images (see Figure 2 and Figure 4). For the ROT ROI the difference between response amplitudes for upright (M = .36, SD = .27) and inverted (M = .13, SD = .13) images was significant (t_{13} = 3.14, p = .008, d = 1.09), while in the MO ROI the difference between upright (M = .21, SD = .18) and inverted (M = .22, SD = .16) images was not significant (t_{13} = .15, p = .882, d = .04). These results are consistent with the results for the photo-realistic images, again supporting the use of the computer generated images as reasonable proxies for faces. They again suggest that the asymmetries in responses for expressions and anti-expressions can partly be attributed to face related processes beyond the early

Accepted manuscript

stages of encoding and further support the selection of the ROT ROI for examining processes related to face encoding.

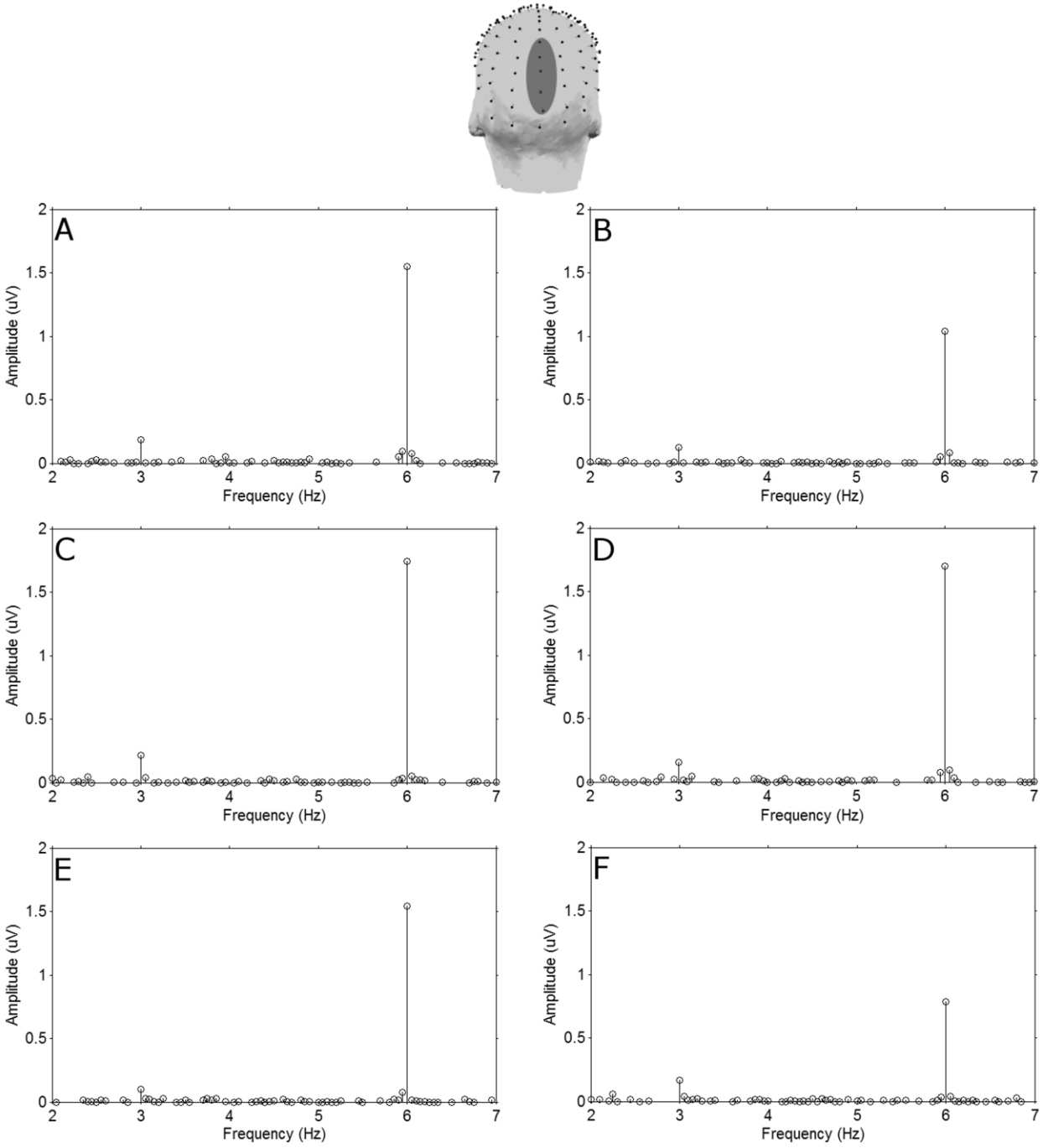


Figure 4. Top scalp map shows the electrodes comprising the MO ROI. (A) Amplitude spectra for the sequence of alternating neutral/expression photo-realistic images, (B) spectra when these images were inverted, (C) spectra for the sequence of neutral/expression computer generated images, (D) the

ACCEPTED MANUSCRIPT

neutral/anti-expression sequence, (E) the expression/anti-expression sequence and (F) spectra when the expression/anti-expression images were inverted.

3.4 Basis for asymmetries

The above analyses of the amplitude spectra address the primary aim of the study and reveal asymmetries in the responses to neutral faces, expressions, and anti-expressions. However, these analyses do not reveal the directions of these asymmetries. To try to evaluate this, we also analyzed the waveforms in the time domain and modelled the phase of the responses in addition to the amplitude. For this, a more conservative Butterworth low-pass filter was applied to the data averaged across trials. This filter comprised a cutoff of 30 Hz, which is more typical of a filter used in ERP studies of face perception (Jacques, d'Arripe, & Rossion, 2007). Data were then segmented into 1 s epochs, resulting in 20 epochs per trial, then averaged within each subject, and finally averaged across the five channels in the ROT ROI. The resulting waveform for one subject in response to the neutral/expression photo-realistic images can be seen in Figure 5. Note that these transformations were only for the purposes of visualizing the recordings and the subsequently reported analyses were performed using the original trial-averaged data described at the end of pre-processing.

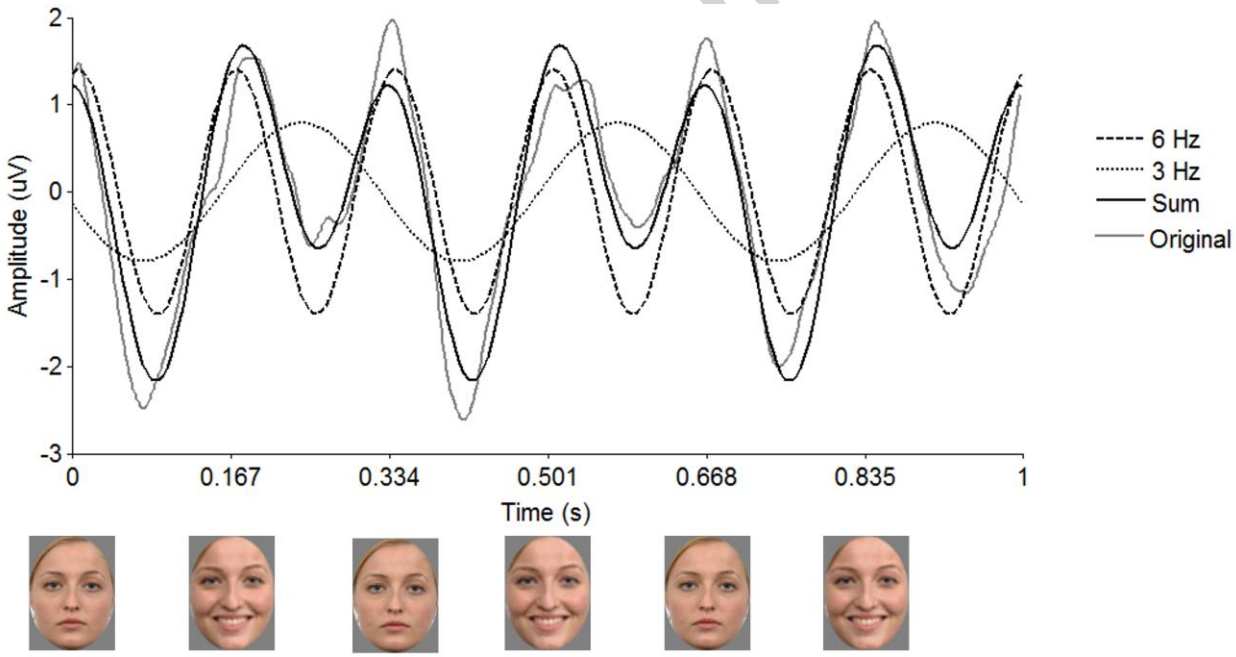


Figure 5. An example of the measured and modeled waveforms for one subject in response to alternations between neutral/expression photo-realistic images. Note that within each condition responses were averaged across the three different expression conditions and both image genders. The female images and expressions shown here are intended to represent the order and onset of image presentation across the time window. Note also that for demonstrative purposes we show here the average of 20 separate time windows of 1 s for the 'original' waveform. When modelling the sum of the

ACCEPTED MANUSCRIPT

3 Hz and 6 Hz responses this was first computed for entire 20 s sequences, then the first and second cycles considered across a 334 ms time window only, as beyond this time point the cycles are repeated.

To quantify differences in amplitude across the original waveforms we extracted the phase and amplitude of the 3 Hz and 6 Hz waves and then fit the sum of these responses for each subject and condition. Fitting the sum of the 3 Hz and 6 Hz waves allows us to simultaneously examine the effect of the 3 Hz wave on the 6 Hz wave in terms of both its phase and amplitude. That is, if the two waveforms are in phase then we will see an amplification of the first cycle of the 6 Hz wave compared to the second cycle, and vice versa if they are out of phase. The amplitude of the 3 Hz wave will determine the magnitude of this amplification. The waves were summed across a 334 ms time window, corresponding to one presentation of the face pair, with each component of the pair defined as a 167 ms time window beginning 50 ms after the onset of an image. The 50 ms delay was included as this may be the earliest time at which responses to faces can be seen (Seeck et al., 1997). A representative example of these fits can be seen in Figure 5 for one subject when observing alternating neutral/expression photo-realistic images. The amplitude of the first cycle of the wave and the second cycle was estimated from the fits for each subject and compared using paired t-tests. For alternations of photo-realistic images of neutral faces and expressions, the amplitude of the first cycle in which a neutral face was presented (M = 1.63, SD = .93) was significantly larger than the amplitude of the second cycle in which an expression was presented (M = 1.24, SD = 1.01) (t_{13} = 3.25, p = .006, d = .4). When these images were inverted the difference in amplitude between the first cycle (M = .83, SD = .71) and the second cycle (M = .95, SD = .83) was not significant (t_{13} = 1.34, p = .204, d = .16), suggesting that the previously observed differences in amplitude are not solely due to low-level properties of the images. For alternations of computer generated images of expressions and neutral faces, the amplitude of the first cycle in which a neutral face was presented (M = 1.15, SD = .72) was not significantly larger than the amplitude of the second cycle in which an expression was presented (M = 1.02, SD = .73) (t_{13} = 1.3, p = .216, d = .18). For alternations between neutral faces and anti-expressions, the amplitude of the first cycle in which a neutral face was presented (M = 1.05, SD = .54) was significantly larger than the amplitude of the second cycle in which an anti-expression was presented (M = .85, SD = .65) (t_{13} = 2.36, p = .034, d = .36). For alternations between expressions and anti-expressions, the amplitude of the first cycle in which an expression was presented (M = 1.39, SD = .73) was significantly larger than the amplitude of the second cycle in which an anti-expression was presented (M = 1.09, SD = .74) (t_{13} = 2.51, p = .026, d = .41). When these images were inverted the difference in amplitude between the first cycle (M = .84, SD = .78) and the second cycle (M = .88, SD = .82) was not significant (t_{13} = .69, p = .502, d = .04), suggesting that the previously observed differences in amplitude are not solely due to low-level properties of the images.

By a simple account of norm-based coding, we might have expected weaker responses to the neutral face. Consistent with this, adaptation to one face prior to an alternation leads to an asymmetry (Retter & Rossion, 2016b), which suggests weaker responses to the adapting face. The present finding that the neutral faces may be tied to the cycle with the larger amplitude is thus surprising. However, results from this analysis should be interpreted with caution. While differences in amplitudes across the waveforms may be accurately quantified, how these differences relate to the specific images is less clear. Unlike ERP studies or other FPVS studies in which there is a larger amount of time in between stimuli of interest (Dzhelyova et al., 2016), presenting facial images at a rate of 6 Hz means that we are measuring many

ACCEPTED MANUSCRIPT

overlapping responses, making it difficult to isolate distinct responses for single images. On the other hand, the present analysis provides additional confirmation that the different face sets led to significant differences in the responses, even if we cannot be confident about the direction of the asymmetry.

3.5 Responses at 6 Hz

While not the main focus of the study a brief examination of responses at the presentation rate of 6 Hz may also further reveal how images in the present study were encoded. Amplitude spectra for the ROT ROI can be seen in Figure 2 and for the MO ROI in Figure 4. Scalp topographies can be seen in Figure 3. Significant responses at the presentation rate of 6 Hz were found in both regions and across all conditions (all $p < .001$) (see Table 2). Continuously significant harmonics with $p < .01$ were observed up to the 3rd harmonic (18 Hz) in all conditions for both ROI's. Baseline-subtracted amplitudes for these harmonics were summed for each condition and region.

Table 2. Z-scores for 6 Hz signal for each condition across the two ROI's.

	Photo-realistic		**Computer generated**			
	Neutral/Expression	**Neutral/Expression (inverted)**	**Neutral/Expression**	**Neutral/Anti-expression**	**Expression/Anti-expression**	**Expression/Anti-expression (inverted)**
MO	82.18	96.54	132.03	116.35	110.48	84.59
ROT	68.67	41.95	60.44	35.39	52.32	36.61

Notes: *All p < .001*

To examine the effect of inversion on responses at 6 Hz, data from the inverted photo-realistic expressions/neutral faces condition and from the inverted computer generated expressions/anti-expressions condition were averaged and compared to the average of responses for these images when upright. In the ROT ROI, response amplitudes for upright faces ($M = 1.66$, $SD = .87$) were significantly larger than for inverted faces ($M = 1.25$, $SD = .88$) ($t_{13} = 3.61$, $p = .003$, $d = .47$). This is to be expected as this region is associated with more face specific processes and inversion disrupts normal face processing. In the MO ROI, mean baseline-subtracted response amplitudes for upright faces ($M = 2.99$, $SD = 1.36$) were also significantly higher than for inverted faces ($M = 1.96$, $SD = .82$) ($t_{13} = 3.61$, $p = .003$, $d = .92$). This is somewhat unexpected as this region is associated more with the processes related to the encoding of low-level properties of the images, which should not be affected by inversion.

To examine differences in the processing of the photo-realistic images and computer generated images, data from the photo-realistic expressions/neutral faces condition were compared to the computer generated expressions/neutral faces condition. For the ROT ROI, response amplitudes were significantly higher for photo-realistic images ($M = 1.78$, $SD = 1.01$) than for computer generated images ($M = 1.48$, $SD = .82$) ($t_{13} = 3.7$, $p = .003$, $d = .33$). This perhaps to be expected as the photo-realistic images are likely better representations of actual faces and may activate face encoding networks to a greater degree than computer generated images. In contrast, for the MO ROI responses for computer generated images ($M = 3.54$, $SD = 1.8$) were significantly higher than for photo-realistic images ($M = 2.96$, $SD = 1.36$) ($t_{13} = 3.71$, $p = .003$, $d = .36$). The larger responses produced by the computer generated images may be due to low-

Accepted manuscript

ACCEPTED MANUSCRIPT

level image statistics not measured in the study, such as greater contrast. This dissociation between the ROI's highlights the different processes captured by the two regions.

4. Discussion

To summarize, we used EEG to examine the neurophysiological responses associated with the processing of neutral faces, expressions and anti-expressions, and the extent to which these processes follow the predictions of norm-based coding. It was reasoned that if expressions and anti-expressions are represented as equivalent deviations from a neutral norm then they should produce equal responses. Images were presented using the FPVS technique (Rossion, 2014a, 2014b), in which two facial images are sequentially presented at a fixed rate (6 Hz) and asymmetries in the responses generated by these images are indicated by the presence of a signal at half the frequency of the presentation rate (3 Hz) (Ales & Norcia, 2009; Retter & Rossion, 2016b). Asymmetries in responses to neutral faces and expressions were observed for photo-realistic images and computer generated images. Asymmetries in responses to neutral faces and anti-expressions as well as expressions and anti-expressions were also observed. Expression/anti-expression images consistently differed in the presence/absence of visible teeth, and could possibly have contributed to the asymmetries in responses. However, inversion of the images significantly reduced these asymmetries, an effect most clearly seen over right occipito-temporal regions, indicating that the reported effects are not solely due to low-level differences between the images and instead may reflect more face-specific process.

4.1 Norm-based coding

Previous studies reporting expression aftereffects following adaptation to anti-expressions have proposed that these faces form opposing pairs within in a norm-based space, similar to an identity and its anti-face (Cook et al., 2011; Rhodes et al., 2017; Skinner & Benton, 2010). However, the asymmetries we observed in the responses generated by these faces suggest that they do not have equivalent "strength" even though they represent equivalent but opposite configural distortions. Of course, had symmetric responses been observed this would not itself demonstrate a norm-based code, since the equivalent amplitudes could have arisen from separate mechanisms for expressions and anti-expressions, rather than equal opposing responses within the same mechanism (in the same way that equal responses might be generated by alternating a vertical and horizontal grating). Conversely, the finding of an asymmetry does point to a tangible difference in how expressions and anti-expressions are encoded. The observed asymmetries could reflect the fact that the anti-expressions are not readily recognizable as a basic expression. As such they may not engage additional processes specific to expression coding or attract the same level of attention as basic expressions (Bekhtereva, Craddock, & Müller, 2015; Wieser, McTeague, & Keil, 2012).

Similar considerations have recently been proposed for the encoding of identity and the effects of familiarity (Faerber, Kaufmann, Leder, Martin, & Schweinberger, 2016). Traditional models of norm-based face coding posit that faces on opposites sides of a norm should be perceived as equally typical. This appears to hold true for unfamiliar faces and their anti-faces. However, familiar faces are perceived

Accepted manuscript

ACCEPTED MANUSCRIPT

as less typical than their anti-faces (Faerber et al., 2016). This finding is particularly interesting when considering that current models of face processing posit that a norm is composed of the average of previously encountered faces and it logically follows that familiar faces are encountered more often than unfamiliar faces. As such, a familiar face should be perceived as more typical yet instead they appear to receive a 'superior representation' (Faerber et al., 2016). Results from the current study similarly imply that a full account of the representation of changeable aspects of faces must take into consideration whether these transformations form familiar and recognizable expressions. The asymmetry in representations of expressions and anti-expressions is further supported by Juricevic and Webster's (2012) finding that the size of aftereffects following adaptation to anti-expressions are generally smaller than following adaptation to expressions, again despite configurally differing from a neutral face by the same degree. Juricevic and Webster (2012) noted that this differs from the face aftereffects for distortions (Rhodes, Jeffery, Watson, Clifford, & Nakayama, 2003; Watson & Clifford, 2003; Webster & MacLin, 1999), or attributes such as gender and ethnicity (Jaquet & Rhodes, 2008; Jaquet, Rhodes, & Hayward, 2007; Little, DeBruine, & Jones, 2005; Ng, Ciaramitaro, Anstis, Boynton, & Fine, 2006; Webster et al., 2004), for which the aftereffects appear more symmetrical.

While the asymmetries between the neural responses for expressions and anti-expressions suggest that these stimuli are not represented as equal and opposite, previous studies of expression aftereffects have nevertheless pointed to norm-based coding, based on findings that these aftereffects increase with the intensity of the expression (Rhodes et al., 2017; Skinner & Benton, 2010). Clues as to how these results might be explained in light of the present findings may be found in earlier studies investigating the distinction between identity and expression encoding. In Bruce and Young's (1986) well-known model of face processing, it is posited that the processing of changeable face properties, such as expression and eye gaze, and invariant properties, such as identity, occur in two separate pathways that bifurcate at the very early stages of encoding (see also Haxby, Hoffman, & Gobbini, 2000). More recently Calder and Young (2005) proposed a refinement of these models in which the functional and neural dissociation of the processing of identity and expression reflects more of a bias in processing in certain neural networks rather than a categorical dissociation, which occurs at a later stage after a common representational system. Possible evidence for this later separation can be found in studies applying principle components analysis (PCA) to sets of facial images varying in identity and expression, a process that extracts elements that account for the largest variation among the stimulus sets. These analyses have shown that while some common components are extracted for expressions and identity there are also many unique components explaining the variance in identity and expressions, indicating that while both aspects of faces can initially be processed within the same system a large degree of dissociation does exist (Calder, Burton, Miller, Young, & Akamatsu, 2001). This process is not proposed as a literal account of the human brain but rather a statistical analogue (Burton, Bruce, & Hancock, 1999) that demonstrates how variant and invariant properties of faces may begin being processed in a common system and then subsequent processing occurring in dissociable routes. In the case of expressions and anti-expressions it may be that they similarly begin by being processed as simple shape distortions, yet only recognizable expressions go on to be processed by neural networks that show biases for specific expressions. The effects of adaptation may partly be occurring at these early and common stages of face encoding before more expression specific encoding. This might also explain how the perception of expressions could be normalized for a neutral expression, despite presumably lopsided exposure to

Accepted manuscript

ACCEPTED MANUSCRIPT

expressions vs. their anti-expressions, by instead normalizing at a level of basic configural coding shared with the representation of face identity (Webster & MacLeod, 2011).

As noted in the Introduction, the possibility of a norm for expressions begs the question of what should be considered a neutral face of an expression. Some studies have approximated the norm by averaging basic expressions (e.g. Rhodes et al., 2017; Skinner & Benton, 2010). In contrast, Cook et al. (2011) constructed their average based on all of the distortions measured in a video recording of an actor reciting jokes. In further contrast, other studies have defined the norm in terms of an actual neutral expression (e.g. Juricevic & Webster, 2012; Rutherford et al., 2008). We followed this last procedure because a neutral expression is arguably a null stimulus for expression percepts and not simply a product of the images chosen by the experimenter. But an appropriate neutral point again depends on what level of coding and type of response is being elicited, and again this remains poorly understood for the visual representation of expressions.

4.2 Basis for asymmetries

In the present study we attempted to gain some insight into the source of the observed asymmetries by further analyzing the phase of the recorded waveforms. This was done by fitting the sum of the 3 Hz and 6 Hz components across a 334 ms time window and comparing the amplitude of the first cycle with the amplitude of the second cycle for each condition. Cycles in which a neutral face was presented were significantly larger in amplitude than cycles in which an expression was presented for photo-realistic images but not computer generated images. Cycles in which neutral faces were presented were also larger in amplitude than cycles for anti-expressions, and cycles for expressions were larger than cycles for anti-expressions. These differences between cycles were reduced when images were inverted, as was the case in the frequency domain analyses.

Presenting facial images at a rate of 6 Hz means that we are measuring many overlapping responses, making it difficult to isolate distinct responses for single images. However, if we assume that the amplitude of a cycle containing the presentation of an image relates more strongly to responses for that image compared to the non-presented image, the data appear to form a pattern of results that may best be explained in terms of typicality of the images. While results from fMRI studies (Loffler et al., 2005) and single cell recordings (Leopold et al., 2006) have shown that larger responses are associated with more distinctive faces, Schulz et al. (2012) have previously shown that in EEG larger amplitudes may be observed for more typical faces compared to more distinctive faces. In the present study we took neutral faces as being representative of the norm and thus can be considered the most typical of all the images used. Expressions are likely the next most typical, followed by anti-expressions, which are likely never seen under normal circumstances. Amplitudes in the current data set followed a similar pattern, with the largest amplitudes observed in cycles containing neutral faces, followed by expressions, and lastly anti-expressions. However, it is important to note that the larger amplitudes for typical faces reported by Schulz et al. (2012) were for variations along identity trajectories and it is unclear whether we should expect expressions and anti-expressions to similarly affect this component. In general the effect of expressions on early EEG components relating to face processing, such as the N170, remains an open question, with some studies finding a difference in amplitudes between expressions and neutral

Accepted manuscript

ACCEPTED MANUSCRIPT

faces (Blau, Maurer, Tottenham, & McCandliss, 2007; Lewis et al., 2003) and others finding no difference (Eimer & Holmes, 2002; Eimer et al., 2003) (for meta-analysis, see Hinojosa, Mercado, & Carretié, 2015). One of the advantages of the FPVS approach is that the frequency-defined responses likely reflect many overlapping ERP responses, thereby avoiding issues of how specific components may respond to different faces. The relatively short intervals between stimulus presentations used in the present study means we cannot make definitive conclusions regarding the relative amplitudes of the responses generated by expression and anti-expressions. However, when longer intervals are used such responses can be examined with more precision (Dzhelyova & Rossion, 2014; Retter & Rossion, 2016a; Rossion et al., 2015). This may prove an interesting avenue for future research and investigations of the norm-based coding of expressions.

4.4 Conclusion

In the present study we used electroencephalography (EEG) to examine the neurophysiological responses associated with the processing of neutral faces, expressions and anti-expressions, and the extent to which these processes follow norm-based coding. Our results point to important asymmetries between expressions and anti-expressions, suggesting that these stimuli do not have equivalent strength of representation at the levels of face coding generating the measured responses. These differences could arise because real expressions are meaningful and engage stronger or additional levels of processing than the configurally equivalent but more ambiguous anti-expressions.

5. Acknowledgments

This work was supported by grants from the National Institutes of Health (EY-10834 to MW; EY-023268 to FJ), with further support for core facilities provided by COBRE P20 GM 103650. The authors would like to thank Talia Retter for her contributions to developing the modelling procedure described in section 3.4, and Adam Sterling for assisting in data collection.

Conflicts of interest: None

Accepted manuscript

ACCEPTED MANUSCRIPT

References

Ales, J. M., & Norcia, A. M. (2009). Assessing direction-specific adaptation using the steady-state visual evoked potential: Results from EEG source imaging. *Journal of Vision*, *9*(7), 1–13. https://doi.org/10.1167/9.7.8

Alvarado, N., & Jameson, K. A. (2002). Varieties of anger: The relation between emotion terms and components of anger expressions. *Motivation and Emotion*, *26*(2), 153–182. https://doi.org/https://doi.org/10.1023/A:1019815402873

Bekhtereva, V., Craddock, M., & Müller, M. M. (2015). Attentional bias to affective faces and complex IAPS images in early visual cortex follows emotional cue extraction. *Neuroimage*, *1112*, 254–266. https://doi.org/10.1016/j.neuroimage.2015.03.052

Benton, C. P. (2009). Effect of photographic negation on face expression aftereffects. *Perception*, *38*, 1267–1275. https://doi.org/10.1068/p6468

Benton, C. P., Etchells, P. J., Porter, G., Clark, A. P., Penton-voak, I. S., & Nikolov, S. G. (2007). Turning the other cheek: The viewpoint dependence of facial expression after-effects. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1622), 2131–2137. https://doi.org/10.1098/rspb.2007.0473

Blanz, V., O'Toole, A. J., Vetter, T., & Wild, H. A. (2000). On the other side of the mean: The perception of dissimilarity in human faces. *Perception*, *29*, 885–891. https://doi.org/10.1068/p2851

Blau, V. C., Maurer, U., Tottenham, N., & McCandliss, B. D. (2007). Behavioral and Brain Functions expression. *Behavioral and Brain Functions*, *3*(7), 1–13. https://doi.org/10.1186/1744-9081-3-7

Blau, V. C., Maurer, U., Tottenham, N., & McCandliss, B. D. (2007). The face-specific N170 component is modulated by emotional facial expression. *Behavioral and Brain Functions*, *3*(1), 1–13. https://doi.org/10.1186/1744-9081-3-7

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 304–327. https://doi.org/https://doi.org/10.1111/j.2044-8295.1986.tb02199.x

Burton, A. M., Bruce, V., & Hancock, P. J. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1–31. https://doi.org/https://doi.org/10.1207/s15516709cog2301_1

Caharel, S., Courtay, N., Bernard, C., Lalonde, R., & Rebaï, M. (2005). Familiarity and emotional expression in influence an early stage of face processing: An electrophysiological study. *Brain and Cognition*, *59*(1), 96–100. https://doi.org/10.1016/j.bandc.2005.05.005

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. https://doi.org/https://doi.org/10.1016/S0042-6989(01)00002-5

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641–652. https://doi.org/10.1038/nrn1724

Calvo, M. G., & Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: An integrative review. *Cognition and Emotion*, *30*(6), 1081–1106. https://doi.org/https://doi.org/10.1080/02699931.2015.1049124

Accepted manuscript

ACCEPTED MANUSCRIPT

Clément, F., Bernard, S., Grandjean, D., & Sander, D. (2013). Emotional expression and vocabulary learning in adults and children. *Cognition & Emotion*, *27*(3), 539–548. https://doi.org/https://doi.org/10.1080/02699931.2012.724012

Cook, R., Matei, M., & Johnston, A. (2011). Exploring expression space: Adaptation to orthogonal and anti-expressions. *Journal of Vision*, *11*(4), 1–9. https://doi.org/10.1167/11.4.2

Dzhelyova, M., Jacques, C., & Rossion, B. (2016). At a single glance: Fast Periodic Visual Stimulation uncovers the spatio-temporal dynamics of brief facial expression changes in the human brain. *Cerebral Cortex*, 1–18. https://doi.org/10.1093/cercor/bhw223

Dzhelyova, M., & Rossion, B. (2014). The effect of parametric stimulus size variation on individual face discrimination indexed by fast periodic visual stimulation. *BMC Neuroscience*, *15*(1), 1–12. https://doi.org/10.1186/1471-2202-15-87

Eimer, M., & Holmes, A. (2002). An ERP study on the time course of emotional face processing. *Neuroreport*, *13*(4), 427–431.

Eimer, M., Holmes, A., & McGlone, F. P. (2003). The role of spatial attention in the processing of facial expression: An ERP study of rapid brain responses to six basic emotions. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(2), 97–110. https://doi.org/https://doi.org/10.3758/CABN.3.2.97

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*, 169–200. https://doi.org/https://doi.org/10.1080/02699939208411068

Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation. *Perception*, *39*(7), 931–941. https://doi.org/https://doi.org/10.1068/p6633

Faerber, S. J., Kaufmann, J. M., Leder, H., Martin, E. M., & Schweinberger, S. R. (2016). The role of Familiarity for Representations in Norm-Based Face Space. *PloS One*, *11*(5), 1–15. https://doi.org/10.1371/journal.pone.0155380

Farah, M. J., & Tanaka, J. W. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 628–634. https://doi.org/10.1037/0096-1523.21.3.628

Fox, C. J., & Barton, J. J. S. (2007). What is adapted in face adaptation? The neural representations of expression in the human visual system. *Brain Research*, *1127*, 80–89. https://doi.org/10.1016/j.brainres.2006.09.104

Freire, A., Lee, K., & Symons, L. A. (2000). The face-inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception*, *29*, 159–170. https://doi.org/10.1068/p3012

Fridlund, A. J. (1997). The new ethology of human facial expressions. In J. A. Russell & J. M. F. Dols (Eds.), *New directions in the study of facial expressions* (pp. 103–129). New York: Cambridge University Press.

Haxby, J. V, Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/https://doi.org/10.1016/S1364-6613(00)01482-0

Heinrich, S. P. (2009). Permutation-based significance tests for multiharmonic steady-state evoked

Accepted manuscript

ACCEPTED MANUSCRIPT

potentials. *IEE Transactions on Biomedical Engineering*, *56*(2), 534–537. https://doi.org/10.1109/TBME.2008.2006021

Hinojosa, J. A., Mercado, F., & Carretié, L. (2015). N170 sensitivity to facial expression: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *55*, 498–509. https://doi.org/https://doi.org/10.1016/j.neubiorev.2015.06.002

Horstmann, G., Lipp, O. V., & Becker, S. I. (2012). Of toothy grins and angry snarls—Open mouth displays contribute to efficiency gains in search for emotional faces. *Journal of Vision*, *12*(5), 1–15. https://doi.org/10.1167/12.5.7

Hsu, S., & Young, A. W. (2004). Adaptation effects in facial expression recognition. *Visual Cognition*, *11*(7), 871–899. https://doi.org/10.1080/13506280444000030

Hurvich, L. M., & Jameson, D. (1957). An opponent-process theory of color vision. *Pscyhological Review*, *64*(6p1), 384–404. https://doi.org/http://dx.doi.org/10.1037/h0041403

Hyvarinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4), 411–430. https://doi.org/https://doi.org/10.1016/S0893-6080(00)00026-5

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review*, *99*(3), 561–565. https://doi.org/10.1037/0033-295X.99.3.561

Jaquet, E., & Rhodes, G. (2008). Face aftereffects indicate dissociable, but not distinct, coding of male and female faces. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 101–112. https://doi.org/10.1037/0096-1523.34.1.101

Jaquet, E., Rhodes, G., & Hayward, W. G. (2007). Opposite aftereffects for Chinese and Caucasian faces are selective for social category information and not just physical face differences. *Quarterly Journal of Experimental Psychology*, *60*(11), 1457–1467. https://doi.org/10.1080/17470210701467870

Jiang, F. J., Blanz, V., & O'Toole, A. J. (2006). Probing the visual representation of faces with adaptation: A view from the other side of the mean. *Psychological Science*, *17*(6), 493–500. https://doi.org/https://doi.org/10.1111/j.1467-9280.2006.01734.x

Juricevic, I., & Webster, M. A. (2012). Selectivity of face aftereffects for expressions and anti-expressions. *Frontiers in Psychology*, *3*, 1–10. https://doi.org/10.3389/fpsyg.2012.00004

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, *68*(1), 1–11. https://doi.org/https://doi.org/10.1016/S0010-0277(98)00035-3

Kaufmann, J. M., & Schweinberger, S. R. (2012). The faces you remember: Caricaturing shape facilitates brain processes reflecting the acquisition of new face representations. *Biological Psychology*, *89*(1), 21–33. https://doi.org/10.1016/j.biopsycho.2011.08.011

Kloth, N., Rhodes, G., & Schweinberger, S. R. (2017). Watching the brain recalibrate: Neural correlates of renormalization during face adaptation. *NeuroImage*, *155*, 1–9. https://doi.org/10.1016/j.neuroimage.2017.04.049

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

Accepted manuscript

ACCEPTED MANUSCRIPT

Leopold, D. A., Bondar, I. V, & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), 3–6. https://doi.org/10.1038/nature04951

Leopold, D. a, O'Toole, a J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89–94. https://doi.org/10.1038/82947

Lewis, S., Thoma, R. J., Lanoue, M. D., Miller, G. A., Heller, W., Edgar, C., … Cañive, J. M. (2003). Visual processing of facial affect. *Neuroreport*, *14*(14), 1841–1845. https://doi.org/10.1097/01.wnr.0000089133.26479.ec

Little, A. C., DeBruine, L. M., & Jones, B. C. (2005). Sex-contingent face after-effects suggest distinct neural populations code male and female faces. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1578), 2283–2287. https://doi.org/10.1098/rspb.2005.3220

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nautre Neuroscience*, *8*(10), 1386–1390. https://doi.org/10.1038/nn1538

Luo, W., Feng, W., He, W., Wang, N. Y., & Luo, Y. J. (2010). Three stages of facial expression processing: ERP study with rapid serial visual presentation. *NeuroImage*, *49*(2), 1857–1867. https://doi.org/https://doi.org/10.1016/j.neuroimage.2009.09.018

N'diaye, K., Sander, D., & Vuilleumier, P. (2009). Self-relevance processing in the human amygdala: gaze direction, facial expression, and emotion intensity. *Emotion*, *9*(6), 798–806. https://doi.org/10.1037/a0017845

Ng, M., Ciaramitaro, V. M., Anstis, S., Boynton, G. M., & Fine, I. (2006). Selectivity for the configural cues that identify the gender, ethnicity, and identity of faces in human cortex. *Proceedings of the National Academy of Sciences*, *103*(51), 19552–19557. https://doi.org/https://doi.org/10.1073/pnas.0605358104

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., & Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, *15*(6), 1–46. https://doi.org/10.1167/15.6.4

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719. https://doi.org/https://doi.org/10.1016/S1388-2457(00)00527-7

Regan, D. (1966). Some characteristics of average steady-state and transient responses evoked by modulated light. *Electroencephalography and Clinical Neurophysiology*, *20*(3), 238–248. https://doi.org/https://doi.org/10.1016/0013-4694(66)90088-5

Retter, T. L., & Rossion, B. (2016a). Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream. *Neuropsychologia*, *91*, 9–28. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2016.07.028

Retter, T. L., & Rossion, B. (2016b). Visual adaptation provides objective electrophysiological evidence of facial identity discrimination. *Cortex*, *80*, 35–50. https://doi.org/10.1016/j.cortex.2015.11.025

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, *46*(18), 2977–2987. https://doi.org/10.1016/j.visres.2006.03.002

Accepted Manuscript

ACCEPTED MANUSCRIPT

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W., & Nakayama, K. (2003). Fitting the mind to the world: Face adaptation and attractiveness aftereffects. *Psychological Science*, *14*(6), 558–566. https://doi.org/https://doi.org/10.1046/j.0956-7976.2003.psci_1465.x

Rhodes, G., & Leopold, D. A. (2011). Adaptive norm-based coding of face identity. In A. J. Calder, G. Rhodes, & J. V Haxby (Eds.), *Oxford Handbook of Face Perception* (pp. 263–286). New York: Oxford University Press.

Rhodes, G., Pond, S., Jeffery, L., Benton, C. P., Skinner, A. L., & Burton, N. (2017). Aftereffects support opponent coding of expression. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 619–628. https://doi.org/http://dx.doi.org/10.1037/xhp0000322

Rhodes, G., Robbins, R., Jaquet, E., McKone, E., Jeffery, L., & Clifford, C. W. (2005). Adaptation and face perception: How aftereffects implicate norm-based coding of faces. In C. W. G. Clifford & G. Rhodes (Eds.), *Fitting the mind to the world: Adaptation and after-effects in high-level vision* (2nd ed., pp. 213–240).

Rossion, B. (2014a). Understanding face perception by means of human electrophysiology. *Trends in Cognitive Sciences*, *18*(6), 310–318. https://doi.org/https://doi.org/10.1016/j.tics.2014.02.013

Rossion, B. (2014b). Understanding individual face discrimination by means of fast periodic visual stimulation. *Experimental Brain Research*, *232*(6), 1599–1621. https://doi.org/10.1007/s00221-014-3934-9

Rossion, B., Delvenne, J., Debatisse, D., & Goffaux, V. (1999). Spatio-temporal localization of the face inversion effect: An event-related potentials study. *Biological Psychology*, *50*(3), 173–189. https://doi.org/https://doi.org/10.1016/S0301-0511(99)00013-7

Rossion, B., Prieto, E. A., Boremanse, A., Kuefner, D., & Belle, G. Van. (2012). A steady-state visual evoked potential approach to individual face perception: Effect of inversion, contrast-reversal and temporal dynamics. *NeuroImage*, *63*(3), 1585–1600. https://doi.org/10.1016/j.neuroimage.2012.08.033

Rossion, B., Torfs, K., Jacques, C., & Liu-Shuang, J. (2015). Fast periodic presentation of natural images reveals a robust face-selective electrophysiological response in the human brain. *Journal of Vision*, *15*(1), 1–18. https://doi.org/10.1167/15.1.18

Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions* (3rd ed., pp. 757–776). New York: Guilford Press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/http://dx.doi.org/10.1037/h0077714

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression ? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102–141. https://doi.org/http://dx.doi.org/10.1037/0033-2909.115.1.102

Russell, J. A., & Fernández-Dols, J. M. (1997). What does a facial expression mean? In J. A. Russell & J.-M. Fernández-Dols (Eds.), *The psychology of facial expression* (pp. 3–30). New York: Cambridge University Press.

Rutherford, M. D., Chattha, H. M., & Krysko, K. M. (2008). The use of aftereffects in the study of relationships among emotion categories. *Journal of Experimental Psychology: Human Perception*

Accepted manuscript

ACCEPTED MANUSCRIPT

and Performance, *34*(1), 27–40. https://doi.org/10.1037/0096-1523.34.1.27

Schulz, C., Kaufmann, J. M., Kurt, A., & Schweinberger, S. R. (2012). Faces forming traces: Neurophysiological correlates of learning naturally distinctive and caricatured faces. *NeuroImage*, *63*(1), 491–500. https://doi.org/10.1016/j.neuroimage.2012.06.080

Schulz, C., Kaufmann, M., Walther, L., & Schweinberger, S. R. (2012). Effects of anticaricaturing vs. caricaturing and their neural correlates elucidate a role of shape for face learning. *Neuropsychologia*, *50*, 2426–2434. https://doi.org/10.1016/j.neuropsychologia.2012.06.013

Schupp, H. T., Arne, O., Jungho, M., Weike, A. I., Stockburger, J., & Hamm, A. O. (2006). The facilitated processing of threatening faces: An ERP Analysis. *Emotion*, *4*(2), 189–200. https://doi.org/10.1037/1528-3542.4.2.189

Seeck, M., Michel, C. M., Mainwaring, N., Cosgrove, R., Blume, H., Ives, J., … Schomer, D. L. (1997). Evidence for rapid face recognition from human scalp and intracranial electrodes. *NeuroReport*, *8*(12), 2749–2754.

Skinner, A. L., & Benton, C. P. (2010). Anti-expression aftereffects reveal prototype-reference coding of facial expression. *Psychological Science*, *21*(9), 1248–1253. https://doi.org/https://doi.org/10.1177/0956797610380702

Skinner, A. L., & Benton, C. P. (2012). The expressions of strangers: Our identity-independent representation of facial expression. *Journal of Vision*, *12*(2), 1–13. https://doi.org/10.1167/12.2.12.Introduction

Srinivasan, R., Russell, D. P., Edelman, G. M., & Tononi, G. (1999). Increased synchronization of neuromagnetic responses during conscious perception. *The Journal of Neuroscience*, *19*(13), 5435–5448. https://doi.org/https://doi.org/10.1523/JNEUROSCI.19-13-05435.1999

Stefanics, G., Csukly, G., Komlósi, S., Czobor, P., & Czigler, I. (2012). Processing of unattended facial emotions: A visual mismatch negativity study. *NeuroImage*, *59*(3), 3042–3049. https://doi.org/10.1016/j.neuroimage.2011.10.041

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *43*(May 2012), 161–204. https://doi.org/10.1080/14640749108400966

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology*, *69*(10), 1996–2019. https://doi.org/https://doi.org/10.1080/17470218.2014.990392

Watson, T. L., & Clifford, C. W. G. (2003). Pulling faces: An investigation of the face-distortion aftereffect. *Perception*, *32*(9), 1109–1116. https://doi.org/https://doi.org/10.1068/p5082

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*(6982), 557–561. https://doi.org/10.1038/nature02420

Webster, M. A., & MacLeod, D. I. A. (2011). Visual adaptation and face perception. *Philosophical Transactions of The Royal Society B*, *366*(1571), 1702–1725. https://doi.org/10.1098/rstb.2010.0360

Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic*

Accepted manuscript

ACCEPTED MANUSCRIPT

Bulletin & Review, *6*(4), 647–653. https://doi.org/https://doi.org/10.3758/BF03212974

Wieser, M. J., McTeague, L. M., & Keil, A. (2012). Competition effects of threatening faces in social anxiety. *Emotion*, *12*(5), 1050–1060. https://doi.org/10.1037/a0027069

Williams, L. M., Palmer, D., Liddell, B. J., Song, L., & Gordon, E. (2006). The 'when'and 'where'of perceiving signals of threat versus non-threat. *NeuroImage*, *31*(1), 458–467. https://doi.org/https://doi.org/10.1016/j.neuroimage.2005.12.009

Highlights

- Configurally equal expressions/anti-expressions produce asymmetric neural responses
- This suggests they not have an equal status in the neural code for expressions
- The plausibility of facial configurations is also important to the representation

Accepted manuscript