*Social Cognitive and Affective Neuroscience*, 2020, 337–346

doi: 10.1093/scan/nsaa043
Advance Access Publication Date: 9 April 2020
Original Article

# An objective and reliable electrophysiological marker for implicit trustworthiness perception

Derek C. Swe,[1] Romina Palermo,[1] O. Scott Gwinn,[2] Gillian Rhodes,[1] Markus Neumann,[1,3] Shanèle Payart,[1] and Clare A. M. Sutherland[1,4]

[1]School of Psychological Science, The University of Western Australia, Crawley, WA 6009, Australia, [2]College of Education, Psychology, and Social Work, Flinders University, Adelaide, SA 5042, Australia, [3]Department of Aviation and Space Psychology, German Aerospace Center (DLR), Hamburg 22335, Germany and [4]School of Psychology, University of Aberdeen, King's College, Aberdeen AB24 3FX, Scotland

Correspondence should be addressed to Derek C. Swe, School of Psychological Science, The University of Western Australia, Crawley, WA, Australia.
E-mail: derek.swe@research.uwa.edu.au

## Abstract

Trustworthiness is assumed to be processed implicitly from faces, despite the fact that the overwhelming majority of research has only involved explicit trustworthiness judgements. To answer the question whether or not trustworthiness processing can be implicit, we apply an electroencephalography fast periodic visual stimulation (FPVS) paradigm, where electrophysiological cortical activity is triggered in synchrony with facial trustworthiness cues, without explicit judgements. Face images were presented at 6 Hz, with facial trustworthiness varying at 1 Hz. Significant responses at 1 Hz were observed, indicating that differences in the trustworthiness of the faces were reflected in the neural signature. These responses were significantly reduced for inverted faces, suggesting that the results are associated with higher order face processing. The neural responses were reliable, and correlated with explicit trustworthiness judgements, suggesting that the technique is capable of picking up on stable individual differences in trustworthiness processing. By demonstrating neural activity associated with implicit trustworthiness judgements, our results contribute to resolving a key theoretical debate. Moreover, our data show that FPVS is a valuable tool to examine face processing at the individual level, with potential application in pre-verbal and clinical populations who struggle with verbalization, understanding or memory.

**Key words:** trustworthiness impressions; EEG; fast periodic visual stimulation; implicit perception; SSVEP

## Introduction

Current theories of trustworthiness perception widely assume that trustworthiness is processed implicitly, that is without needing explicit impression formation instructions (Willis and Todorov, 2006; Krumhuber *et al.*, 2007; Van't Wout and Sanfey, 2008; Walker and Vetter, 2016; Zebrowitz, 2017). This assumption seems intuitively plausible given the nature of trustworthiness perception in everyday life: we form split-second evaluations of trustworthiness when meeting new people (Willis and Todorov, 2006) and the trustworthiness of a face has been found to affect court rulings, financial lending and even leadership choices (Olivola and Todorov, 2010; Olivola *et al.*, 2014). However, current theories of trustworthiness perception are almost entirely founded on research that explicitly requires individuals to form impressions (see Todorov *et al.*, 2015, for a review). Moreover, it has recently been questioned whether trustworthiness is perceived implicitly or not

**Received:** 21 October 2019; **Revised:** 18 February 2020; **Accepted:** 18 March 2020

(Winston *et al.*, 2002; Santos and Young, 2005; Klapper *et al.*, 2016). A demonstration of implicit processing of trustworthiness would therefore have important implications for current theory: if implicit trustworthiness can be demonstrated, these widespread assumptions are validated. If not, or if implicit and explicit trustworthiness processing diverges, then this important theoretical gap needs to be accounted for.

Behavioral investigations of memory for faces have found evidence for and against the implicit encoding of trustworthiness. Klapper *et al.* (2016) used the 'who said what' paradigm (Taylor *et al.*, 1978) to test whether facial trustworthiness was implicitly encoded in memory. This paradigm measures the extent to which people use facial (or other) cues to remember who said a particular statement, thereby measuring whether that cue was potentially spontaneously encoded in memory. People were more likely to misattribute statements made by trustworthy-looking speakers to other trustworthy-looking speakers, rather than to untrustworthy-looking speakers. This finding suggests that facial trustworthiness was implicitly coded and was used to categorize individuals (Klapper *et al.*, 2016), although participants could have also been using trustworthiness cues strategically, to aid memory. In contrast, Santos and Young (2005) did not find evidence for implicit trustworthiness perception in an elegant study based on the 'isolation effect' (Von Restorff, 1933; Hunt, 1995). The isolation effect reflects better memory for items (e.g. female faces) surrounded by those from a different category instead of the same category (e.g. male instead of female faces). Isolation effects were found for age and gender, but not for trustworthiness, indicating that trustworthiness was not perceived implicitly (Santos and Young, 2005).

A few event-related potential (ERP) studies have now investigated the time course and temporal dynamics of trustworthiness processing (Yang *et al.*, 2011; Dzhelyova *et al.*, 2012). However, the ERP components involved in trustworthiness processing are not currently well defined (Yang *et al.*, 2011). In addition, most ERP studies have involved explicit trustworthiness ratings, with two notable exceptions. Lischke and collegues (2018) asked participants to view trustworthy and untrustworthy faces without judging trustworthiness. Larger amplitudes were found in late positive potentials over centro-parietal areas for untrustworthy than trustworthy faces, suggesting implicit trustworthiness processing at 500–800 ms after face onset (Lischke *et al.*, 2018). Given that the effects were observed at a relatively late ERP component, they could potentially reflect semantic rather than early visual processing (Schweinberger and Neumann, 2016). Dzhelyova *et al.* (2012) also recorded ERPs while participants judged either the trustworthiness (explicit condition) or gender (implicit condition) of faces. Implicit processing was found at the level of the face-sensitive N170 over occipito-temporal areas, but only for female trustworthy and male untrustworthy faces. In contrast, effects of explicit, but not implicit, trustworthiness processing were found in the early posterior negativity at 230–280 ms after face onset (Dzhelyova *et al.*, 2012). Therefore, there is some evidence for implicit coding of trustworthiness from ERP research, but ERP components of implicit trustworthiness perception appear somewhat inconsistent across paradigms.

Finally, functional magnetic resonance imaging (fMRI) has shown that different blood-oxygen-level-dependent (BOLD) signals are elicited by faces that vary in trustworthiness, even in the absence of explicit trustworthiness judgements (Winston *et al.*, 2002; Engell *et al.*, 2007; Todorov *et al.*, 2011; Mattavelli *et al.*, 2012). Differences have primarily been observed in the amygdala, which possibly suggests affective processes beyond initial visual encoding (Winston *et al.*, 2002; Engell *et al.*, 2007). Interestingly, more recent studies have also observed BOLD signal differences in response to face trustworthiness in the occipital visual cortex (Todorov *et al.*, 2011; Mattavelli *et al.*, 2012). These latter results are particularly striking, as they suggest that implicit trustworthiness processing may occur not only in the amygdala, as previous research has shown, but also in areas traditionally considered to be basic face processing regions (Haxby *et al.*, 2000). What remains unclear is the precise relationship among these areas and the potential pathways and directions along which trustworthiness information is passed.

To summarize, findings from behavioral, ERP and neuroimaging paradigms are conflicting, leading to an unresolved debate as to whether trustworthiness is processed implicitly (Winston *et al.*, 2002; Santos and Young, 2005; Klapper *et al.*, 2016). Moreover, no clear ERP component has been consistently associated with trustworthiness processing. One problem for the field is the lack of a reliable, objective marker of face trustworthiness processing. Developing such a marker would greatly help to address this current theoretical debate. It would allow the testing of populations that struggle to verbalize, understand or remember trustworthiness judgements, which is important given the rising interest in understanding trustworthiness processing in children and other special populations (Ewing *et al.*, 2019; Mondloch *et al.*, 2019; Sutherland *et al.*, 2019).

A potential solution lies with a technique, fast periodic visual stimulation (FPVS), that has recently been used to great effect in investigating implicit face perception, such as discriminating facial identity (Rossion, 2014). FPVS is used in conjunction with electroencephalography (EEG) frequency analyses of the steady-state visual evoked potential (Regan, 1966). In a face perception FVPS paradigm, face images are rapidly presented at a predetermined rate (e.g. 6 images/s, resulting in a frequency of 6 Hz). Within this sequence, a sub-group of images that differ from the others on a certain attribute (e.g. identity) is presented at a different, oddball rate (e.g. once per second/every sixth face, resulting in a frequency of 1 Hz). If the brain is sensitive to the attribute varying at the oddball frequency, then a peak in amplitude will be present in the EEG signal at that frequency. The FPVS paradigm has recently been successfully used to investigate the neural discrimination of facial identity (Rossion, 2014), facial emotion (Dzhelyova *et al.*, 2016; Zhu *et al.*, 2016; Gwinn *et al.*, 2018) and, more recently, higher level socio-cognitive processes such as attractiveness discrimination (Luo *et al.*, 2019), visual perspective taking (Beck *et al.*, 2017) and semantic categorization (Stothart *et al.*, 2017).

Critically, during FPVS, processing is measured concurrently, and there is need for explicit instructions neither to judge the faces nor to hold faces in memory. As FPVS involves frequency tagging, responses occur at a predefined frequency without requiring knowledge of the specific processing components involved (Liu-Shuang *et al.*, 2014; Rossion, 2014). In this way, FPVS has a high degree of objectivity and specificity. The frequency response that results from FPVS has a high signal-to-noise ratio, making it highly reliable (Liu-Shuang *et al.*, 2014). Finally, another benefit of FPVS is that it can be applied to examine neural responses at the individual level, without the expense of neuroimaging. Taken together, these characteristics make FPVS an ideal technique for understanding implicit face perception. Here, we use FPVS to examine whether facial trustworthiness perception can occur implicitly at the group and individual level.

### The present study

The primary aim of this study is to use FVPS for the first time to investigate whether facial trustworthiness can be processed implicitly. To do this, we utilize an oddball FPVS paradigm, inspired by recent studies on facial identity (Rossion, 2014) and facial emotion (Zhu *et al.*, 2016). We focus on measuring whether trustworthiness processing can occur in areas associated with face perception (e.g. occipito-temporal cortex: Haxby *et al.*, 2000), given the recent fMRI literature suggesting that occipito-temporal cortex may subserve these higher order judgements (Todorov *et al.*, 2011; Mattavelli *et al.*, 2012). We expect that the visual cortex will be sensitive to visual changes in trustworthiness, which should induce an amplitude spike in neural activity at the trustworthiness oddball frequency and its harmonics.

A secondary aim is to assess the potential for the FVPS technique to be used in the future as a neural marker of trustworthiness processing at the individual level. If the technique can capture stable individual responses, it could be used in developmental, clinical or individual differences research, which is important given the rising interest in understanding trustworthiness processing in these fields (Hehman *et al.*, 2017; Ewing *et al.*, 2019; Mondloch *et al.*, 2019; Sutherland *et al.*, 2019). To this end, we investigate whether there are any stable individual differences in the neural discrimination response to facial trustworthiness, by measuring test–retest reliability. Finally, we also test whether neural trustworthiness discrimination correlates with explicit impressions of trustworthiness.

## Materials and methods

### Participants

The final sample consisted of 31 participants (14 males, ages ranging from 18 to 42, $M = 24$ years, s.d. = 6 years). Sample size was based on a power analysis (conducted in R, version 3.6.1). We used the effect size ($d = 1.16$) from a conceptually related face perception oddball FPVS study (Beck *et al.*, 2017). The power analysis showed that 28 participants were needed to find this effect size with 0.99 power for $P \leq 0.05$. Participants were students at the University of Western Australia ($N = 14$) or recruited from the wider community ($N = 17$). Only Caucasian participants were tested to control for potential other-race effects (Hancock and Rhodes, 2008), as Caucasian face stimuli were used. The study was approved by the University of Western Australia human ethics committee.

### Stimuli

Twenty pairs of faces, each consisting of a trustworthy and an untrustworthy version of the same face identity, were taken from a database of computer-generated FaceGen images (for original trustworthiness ratings, see Todorov *et al.*, 2013). Faces were forward-facing with direct gaze. Only male-looking faces were chosen, as previous research has shown interaction effects between gender and trustworthiness of faces (Dzhelyova *et al.*, 2012; Sutherland *et al.*, 2015). Trustworthy faces had a level of +1 s.d. on the trustworthiness dimension, while untrustworthy faces had a level of −3 s.d. This asymmetry was as a precaution, due to the concern that increasing trustworthiness further made the male faces start to look androgynous or female.

Faces were converted to greyscale, and then luminance and contrast adjusted to the average of all images using the SHINE toolbox (Willenbockel *et al.*, 2010) in MATLAB R2017b (The MathWorks, USA), to control for low-level features of the face images driving the neural responses (Figure 1). Additionally, face image size was jittered across presentations within sequences (ranging from 80 to 120% in 2% steps in each sequence) to avoid low-level adaptation.

### Procedure

An FPVS oddball paradigm was used (Liu-Shuang *et al.*, 2014; see Figure 1). Five trustworthy faces were presented in sequence, followed by an untrustworthy oddball face (or vice versa) so that facial trustworthiness changed at 1 Hz within a base rate of 6 Hz. All 20 face identities were shown equally often as base and oddball faces across the different sequences to avoid trustworthiness being confounded with identity. Thus, there were five trustworthy oddball sequences and five untrustworthy oddball sequences, each consisting of 20 face identities as base images (e.g. trustworthy or untrustworthy) and four face identities as oddball images (e.g. untrustworthy or trustworthy, respectively). Each sequence lasted 40 s, consisting of 240 faces (10 repetitions of the 20 base identity faces and 10 repetitions of the four oddball faces, with a different set of faces shown depending on the trial). An inverted condition was also included, which was identical to the upright condition but with the faces rotated 180°. Inverting a face disrupts face-selective processing while keeping the processing of low-level features intact (Rhodes *et al.*, 1993; Rossion and Gauthier, 2002; Yovel and Kanwisher, 2005). Critically, comparing the inverted condition with the upright condition serves to test whether the neural response was face-selective, while also controlling for low-level visual influences (e.g. local luminance, contrast, spatial frequency, angle or curvature). When testing individual differences, the inverted condition was used as a baseline to avoid confounding differences in general neural responsiveness, extraneous electrical noise, and so on. In total, there were 10 upright sequences and 10 inverted sequences.

Faces were presented using a square wave function with a 100% duty cycle. That is, each face was shown at full contrast for the full duration of each cycle of the square wave (167 ms, i.e. 1000 ms/6 faces), with the next face appearing immediately. To help maintain attention throughout each trial, participants were asked to press a key whenever the central fixation cross changed to a square, which happened eight times during each sequence at random time points. Importantly, participants were only instructed to complete the fixation task and look at the faces. No judgements relating to trustworthiness, or otherwise, were asked in response to the faces during the FPVS task. The FPVS task was run using a custom java-based program.

In order to measure test–retest reliability, we repeated the entire task across a second, identical, block and calculated the correlation between neural trustworthiness discrimination responses across two blocks across the participants. To allow us to best capture any individual differences, face sequences were shown in the same pseudorandom order across all participants.

After the FPVS tasks were completed, participants were asked to explicitly rate the trustworthiness of the faces on a scale of 1–9 (1 = not at all trustworthy to 9 = extremely trustworthy) using Qualtrics (Provo, UT). The rating task served as a manipulation check and allowed us to compare neural responses with explicit judgements of trustworthiness. The experiment took approximately 60 min, including 15 min to set up the EEG.

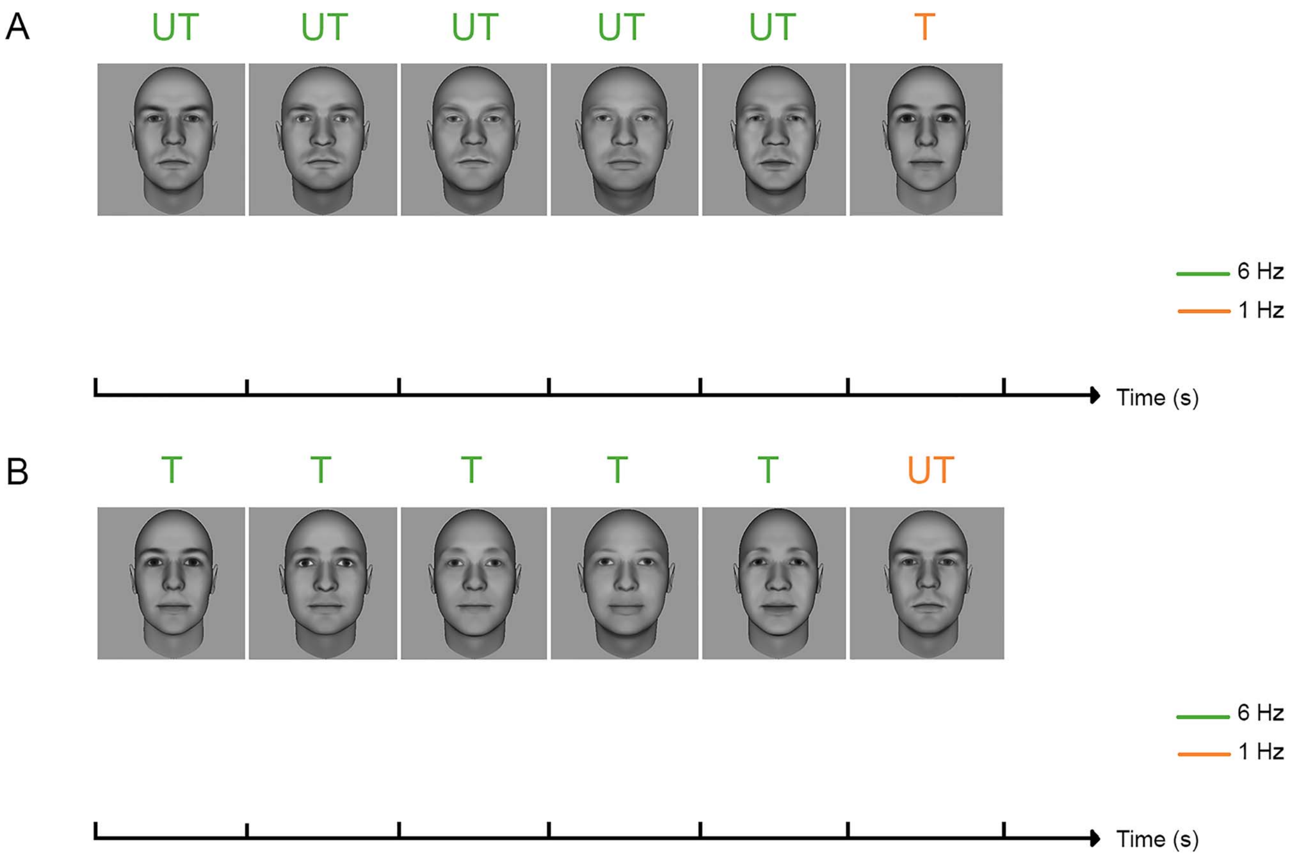

**Fig. 1.** Example (A) trustworthy and (B) untrustworthy oddball FPVS sequences. Images were shown at a frequency of 6 Hz, with oddball images shown at a frequency of 1 Hz. Thus, the sequence of five untrustworthy base (UT) face images was followed by one trustworthy oddball (T) (and vice versa for trustworthy oddball sequences). Faces were shown using a square wave function with a 100% duty cycle such that the next face appeared as soon as the previous face disappeared. There were equal numbers of trustworthy and untrustworthy oddball sequences in each block of the task.

## EEG analyses

***EEG acquisition.*** EEG data were recorded using a 64-channel Biosemi ActiveTwo system (Biosemi, Amsterdam, Netherlands), using the extended 10–20 layout (see http://biosemi.com/pics/cap_64_layout_medium.jpg). The Biosemi DRL/CMS circuit was used as the recording reference (http://biosemi.com/faq/cms&drl.htm). Electrode offsets were kept below 30 mV. The EEG recording was digitized at 2048 Hz and then down-sampled to a rate of 512 Hz.

***EEG pre-processing.*** The EEG recordings were analyzed using Letswave 6 (http://www.nocions.org/letswave6) running over MATLAB 2017b (MathWorks, USA). Standard FPVS processing procedures were followed (Rossion *et al.*, 2012; Retter and Rossion, 2016; Gwinn *et al.*, 2018). EEG waveforms from both blocks were merged for the main analyses. Each block was analyzed separately for test–retest reliability analysis. The EEG data were initially bandpass-filtered at a high-pass cut-off of 0.1 Hz and a low-pass cut-off of 120 Hz, using a Butterworth filter (order 4). Electrical line noise was also filtered out at 50 Hz plus two harmonics with a fast Fourier transform multi-notch filter. Data were then segmented with an additional 2 s baseline before and after the sequences (−2 to 42 s). Four participants were identified who on average blinked more than 0.2 times/s during the 40 s stimulation sequences, using a blink detection plugin for Letswave. This criterion was chosen based on previous FPVS studies (Retter and Rossion, 2016; Gwinn *et al.*, 2018). For these individuals, blink corrections were applied using an independent component analysis with a square matrix (Retter and Rossion, 2016). Across all participants, three instances of excessively noisy channels (channels with amplitude deviations greater than 200 μV) occurred and were replaced with the average of three neighboring channels using interpolation. No more than one channel was interpolated per participant.

All 64 channels were re-referenced to the average of all 64 electrodes and then re-segmented to the last presentation of the oddball (i.e. 40 s). Initial analyses showed similar responses regardless of oddball type. Thus, waveforms were averaged across both trustworthy and untrustworthy oddball conditions to improve the SNR of the recordings. The data were then averaged across trials, within the upright and inverted conditions, resulting in two waveforms for each subject (or four waveforms when looking at each block separately). These waveforms were then transformed to the frequency domain using a fast Fourier transform filter. When comparing responses across participants and conditions, baseline corrections were applied. This correction took the form of a baseline subtraction in which the average of the 20 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) (Rossion *et al.*, 2012). This procedure was carried out to control for differences in baseline noise across participants and across the frequency spectrum within participants. When determining the significance of frequency-locked responses, z-scores were calculated [z = (x-baseline)/standard

**Table 1.** Z-scores for the fundamental frequency and harmonics (up to the eighth harmonic) at the ROI (electrodes P8, PO8 and P10)

| | 1 Hz | 2 Hz | 3 Hz | 4 Hz | 5 Hz | 7 Hz | 8 Hz |
|---|---|---|---|---|---|---|---|
| Upright | 1.65* | 3.98*** | 8.53*** | 5.75*** | 6.00*** | 2.63*** | 2.30** |
| Inverted | -0.39 | 2.45** | 0.42 | 0.27 | 1.42 | 1.97* | 2.36** |

Note: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$ (one-tailed).

deviation of the baseline]. Baselines were defined using the same bin range as the baseline subtraction described above, with the exception that the minimum and maximum amplitude bins were now also included to provide a more conservative test of significance (Rossion *et al.*, 2012). We focused our analysis on a pre-defined right occipito-temporal region of interest (ROI), comprising electrodes over scalp regions previously shown to be associated with face processing (electrodes P8, P10 and PO8: Dzhelyova and Rossion, 2014; Retter and Rossion, 2016).

When performing group-level analyses, the SNR spectra were grand-averaged across participants, separately for upright and inverted conditions. For significance testing, z-scores were computed from the amplitude spectra grand-averaged across participants, separately for each condition. When comparing amplitudes, the sum of the baseline-subtracted harmonics (including the fundamental 1 Hz oddball frequency) was used. Harmonics up to and including the last significant consecutive harmonic (the eight harmonic, i.e. 8 Hz) were summed for both upright and inverted conditions. We excluded the 6 Hz presentation frequency as this frequency represents the generic face detection response.

## Results

### Manipulation and attention check

As expected, the trustworthiness face manipulation was successful. The trustworthy faces were explicitly rated as significantly more trustworthy ($M = 6.56$, $s.d. = 0.5$) than the untrustworthy faces ($M = 3.9$, $s.d. = 0.4$), $t(30) = 2.41$, $P < 0.001$, $d = 5.82$.

Participants also passed the attention check, with a mean detection rate of fixation cross changes of 99% and a minimum of 93% across participants.

### A neural discrimination response to facial trustworthiness

If individuals are able to perceive trustworthiness implicitly, they should show a neural discrimination response to facial trustworthiness, indicated by signals at the fundamental oddball frequency and its harmonics. There was no significant difference in the strength of the neural response between the trustworthy and untrustworthy oddball sequences in the upright condition: $t(30) = -0.67$, $P = 0.51$. Therefore, these two conditions were collapsed to form one trustworthiness oddball signal during the precise segmentation of the epochs, which was used in further analyses. Critically, the fundamental frequency (1 Hz) and each of the harmonics (2, 3, 4, 5, 7 and 8 Hz) were significant in the upright condition (Table 1), suggesting that facial trustworthiness can be processed implicitly.

We then calculated the absolute strength of the oddball signal, by taking the sum of the fundamental frequency and its significant harmonics. This summed oddball response represents the strength of the neural response to changes in facial trustworthiness. The summed trustworthiness oddball response was used to compare upright and inverted faces and for analyses at the individual participant level. The summed oddball response in the upright condition was significant ($z = 7.24$, $P < 0.001$), also indicative of implicit processing of trustworthiness. Consistent with the absence of an interaction with the trustworthiness condition, the summed oddball was also significant across both the untrustworthy ($z = 4.26$, $P < 0.001$) and trustworthy ($z = 3.27$, $P < 0.001$) upright oddball conditions.

To demonstrate that the trustworthiness discrimination response for upright faces was not due to low-level visual differences of the stimuli, we examined the inverted face condition. Importantly, the summed oddball response was not significant in the inverted condition, $z = 1.30$, $P = 0.97$, nor was there a consistent pattern of significant amplitude spikes in the inverted condition at individual frequencies of interest (Table 1, also see Supplementary Figure S2). Indeed, the summed oddball response was significantly higher in the upright ($M = 0.28$, $s.d. = 0.31$) compared with the inverted ($M = 0.07$, $s.d. = 0.16$) condition across participants, $t(30) = 3.78$, $P < 0.001$, indicating that low-level visual stimuli differences cannot account for the trustworthiness neural discrimination response. Oddball responses and scalp topographies for both upright and inverted conditions are shown in Figure 2.

In addition to the right occipito-temporal ROI chosen as our focal analysis due to its importance in face processing, several other regions also had significant responses to trustworthiness (Figure 2). These channels were either close to the ROI, located near the base of the scalp or located over the left occipito-temporal face processing areas. Tentatively, this pattern may suggest that the processing of trustworthiness is not localized to the right face processing areas but extends to the left-hemispheric face processing areas and potentially other visual areas of the brain as well. However, inspection of the scalp topographies indicates that the strongest activity is found over the right occipito-temporal region.

### Individual neural discrimination responses

The secondary aim of the study was to explore the potential to use FPVS to identify a reliable neural marker for facial trustworthiness perception at the individual level. Thus, we calculated test–retest reliability at the individual participant level by correlating the oddball response in the upright condition between the first and second FVPS blocks, which were identical in terms of stimulus presentation (Figure 3 shows the individual scalp topographies for each block). Test–retest reliability was reasonably good: $r(30) = 0.499$, $P = 0.004$, $N = 31$, indicating that the trustworthiness neural discrimination response is a moderately stable construct. To control for individual factors and low-level differences between stimuli, we also calculated the reliability of the upright condition after residualising for the inverted condition (DeGutis *et al.*, 2013). That is, we ran a linear regression using the upright condition responses as the dependent variable and the inverted condition responses as the independent variable. The residuals obtained from this procedure measure the variance in the upright condition after controlling for the control (inverted) condition. We correlated these residuals across blocks, which showed moderate test–retest reliability ($r = 0.303$, $P = 0.098$, $N = 31$).

Additionally, we also calculated internal reliability by splitting the sequences in half. Five sequences were randomly selected, with three sequences belonging to one oddball condition and two belonging to the other (i.e. three trustworthy and two untrustworthy oddball sequences correlated against two trustworthy and three untrustworthy oddball sequences). This method also showed a good reliability ($r = 0.67$, $P < 0.0001$, $N = 31$). Using residuals based on regressing out the inverted condition responses from the upright condition responses also showed a good internal reliability ($r = 0.67$, $P < 0.0001$, $N = 31$).

We also ascertained that there was individual variation in the neural oddball response to trust (see Supplementary Table S1 for the z-scores of each individual participant for upright and inverted faces, which for upright faces range from −1.04 to 7.82), with largest responses to upright faces primarily localized to the right occipito-temporal region (Figure 3). We then calculated the average trustworthiness rating from the face rating task for each participant and correlated these average ratings with the neural oddball response across participants. We found a significant negative correlation between the explicit trustworthiness ratings for the face stimuli, and the neural trustworthiness discrimination response (after residualizing for the inverted condition: $r = -0.41$, $P = 0.023$). Neural responses were stronger (i.e. discrimination was greater) for participants who rated faces as less trustworthy overall. There were no other significant correlations between explicit trustworthiness ratings (ratings of high trust faces, low trust faces or their difference) and the neural trustworthiness discrimination response (all $r < -0.34$ all $P > 0.06$).

## Discussion

We investigated whether facial trustworthiness can be perceived implicitly, applying a novel methodology to this question in the form of an FPVS paradigm in conjunction with EEG. Significant neural responses were observed at frequencies corresponding to a visual change in trustworthiness, indicating that trustworthiness can be perceived implicitly. The neural discrimination response was significantly reduced and became non-significant when faces were inverted, indicating that the response represented high-level face processing, and was not solely due to low-level visual differences between the stimuli. Furthermore, we found that the neural trustworthiness discrimination response was reliable, and found preliminary evidence to suggest that the response was associated with explicit trustworthiness ratings. Crucially, our results help address a recent theoretical debate (Winston *et al.*, 2002; Santos and Young, 2005; Klapper *et al.*, 2016) by demonstrating that face trustworthiness can be processed implicitly. Moreover, our study contributes to methodology by demonstrating that this technique can identify individual implicit trustworthiness perception, which will aid future clinical, developmental and individual differences research.

Our findings provide an important source of converging evidence with behavioral, EEG and fMRI studies that have found evidence for implicit processing of trustworthiness (Winston *et al.*, 2002; Engell *et al.*, 2007; Klapper *et al.*, 2016; Lischke *et al.*, 2018). However, our results vary from those of Santos and Young (2005), who did not find evidence for implicit trustworthiness perception with an isolation effect memory paradigm. Our FPVS paradigm differs from previously used designs; in that, it may be particularly sensitive to implicit processing, as it does not rely on memory, and can measure aspects of face perception objectively by pre-specifying where in the neural signal, the response should occur (Rossion *et al.*, 2012). Therefore, FPVS does not require the prior knowledge of the specific ERP processing components involved in trustworthiness perception, which are currently not well defined (Yang *et al.*, 2011). Importantly, our results particularly support recent findings that face selective brain regions are also involved in trustworthiness processing (Todorov *et al.*, 2011; Mattavelli *et al.*, 2012), as our neural trustworthiness discrimination response was localized over areas typically understood to be associated with face processing.

The secondary aim of our study was to determine whether FPVS could be applied to identify a novel neural marker for trustworthiness perception. A recent study has shown surprisingly high disagreement between individuals in their perceptions of social traits including trustworthiness, finding that as much as half of the variance in facial trustworthiness impressions is due to unique differences specific to the observer (Hehman *et al.*, 2017). However, individual differences in trustworthiness perception are still poorly understood. Critically, we did find that the FPVS measure was reliable and presented good variability across participants, allowing future research to use this technique to test individual differences in trustworthiness processing. Compared with fMRI, EEG is relatively inexpensive and more accessible, making FPVS ideal for large-scale individual difference studies.

We also found that participants who rated faces as less trustworthy showed stronger neural trustworthiness discrimination. People in general are more sensitive to differences in untrustworthy faces than perceptually equivalent differences in trustworthy faces, suggesting that untrustworthiness is especially salient, likely due to its relevance for threat detection (Oosterhof and Todorov, 2008). Moreover, individuals who are more chronically untrusting (i.e. sensitive to being victimized by others; Schmitt *et al.*, 2010) are more sensitive to untrustworthiness cues (Gollwitzer *et al.*, 2013). Thus, people who are chronically less trusting will find more faces to be untrustworthy and may therefore also show stronger neural trustworthiness discrimination. Future research could confirm this idea by relating neural trustworthiness discrimination to individual differences in chronic trusting disposition.

Finally, our results also validate the future use of FVPS for testing trust perception in populations that would otherwise prove difficult to examine due to the limitations of typical verbal paradigms. We have shown that the FPVS task can track trustworthiness perception without requiring any verbal responses or semantic understanding of trustworthiness, making it an ideal method of testing preverbal children, or clinical populations who show difficulties in verbal or cognitive abilities. This methodological advance is important, as there has been growing interest in tracking the development of trust perception (Ewing *et al.*, 2019; Mondloch *et al.*, 2019), as well as clinical differences (Sutherland *et al.*, 2019).

### Future research directions

The use of FVPS to investigate high-level social attributes is an emerging field of research (Beck *et al.*, 2017; Luo *et al.*, 2019), and our results further establish the robust applicability of FPVS for higher level social face attributes such as trustworthiness. Importantly, there are many other higher level face attributes and social traits that could be investigated with FVPS. For example, future research could apply the FPVS paradigm to test the implicit perception of other key face dimensions taken

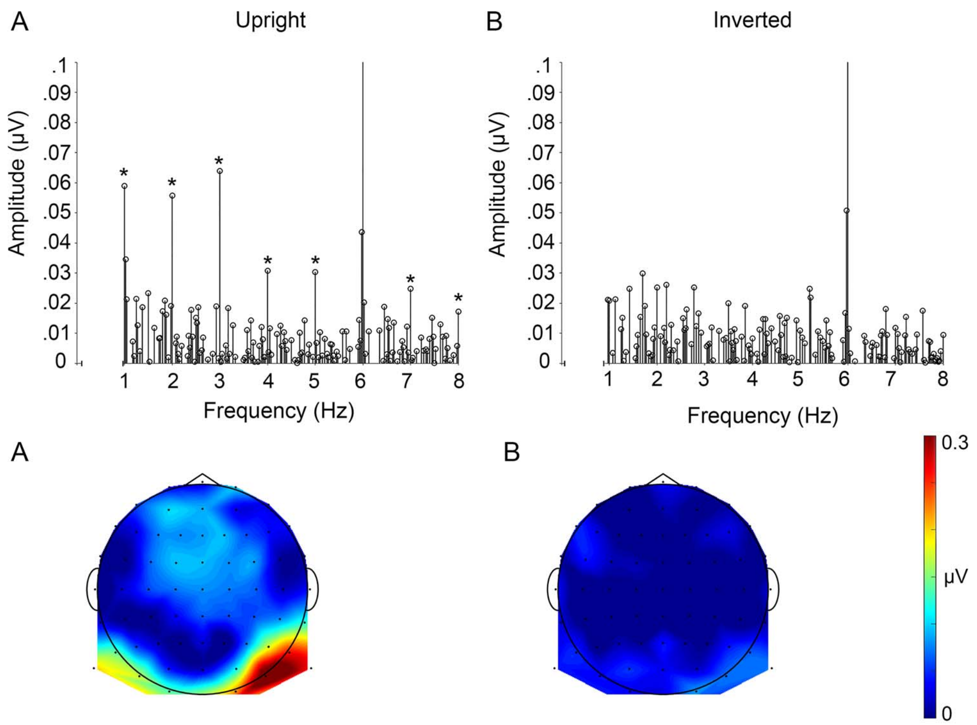

**Fig. 2.** Oddball response amplitude spectra and scalp topographies for the (A) upright and (B) inverted conditions. Top row: baseline subtracted amplitude spectra, collapsed across both trustworthy and untrustworthy oddball face stimuli at the right occipito-temporal ROI. This ROI consists of electrodes P8, PO8 and P10 (z-scores for each electrode are in Supplementary Table S2). Bottom row: scalp topographies for the overall trustworthiness oddball response (sum of fundamental oddball frequency and its significant harmonics), grand averaged across participants (see Supplementary Figure S1 for topographies for each harmonic). Note that the 6 Hz response reflects the generic face detection response rather than the trustworthiness discrimination response and is not included in the sum of harmonics. $* P < 0.05$. All z-value tests report one-tailed $P$ values, which are appropriate here as the signal is only ever meaningful above zero.

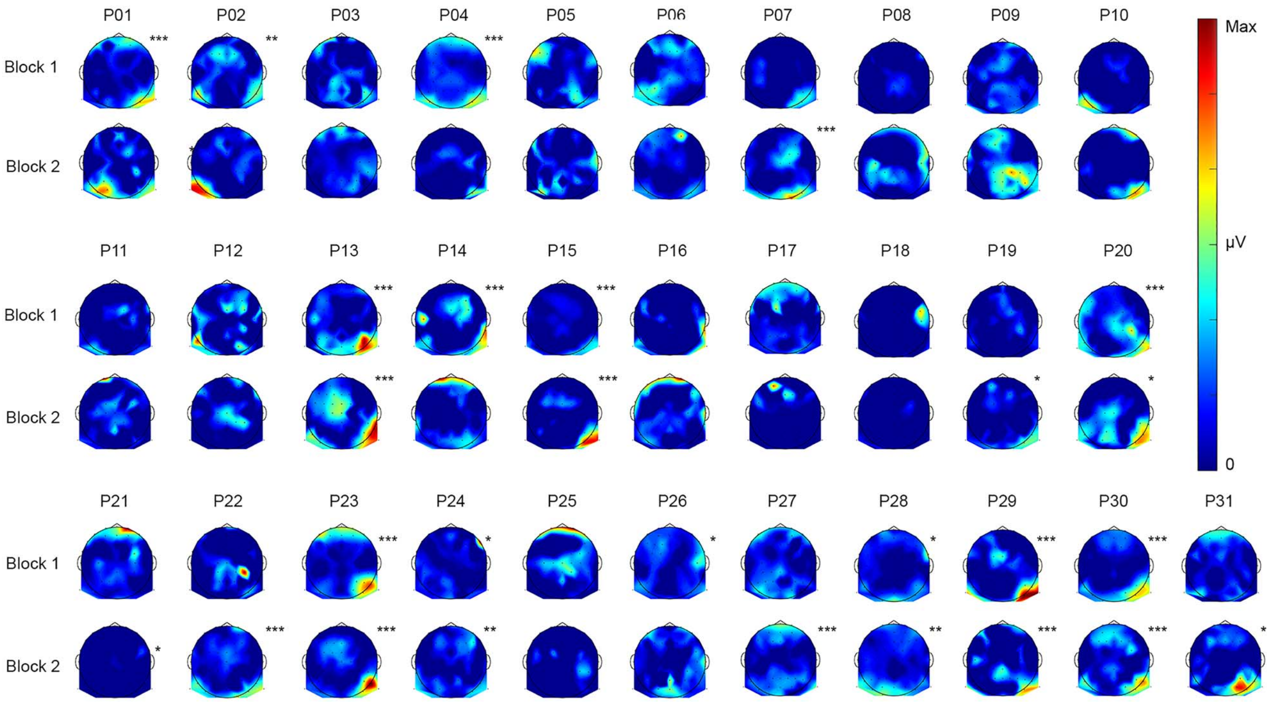

**Fig. 3.** Scalp topographies of the normalized sum of the harmonics for each participant (McCarthy and Wood, 1985). Of each set, the top row represents Block 1 and the bottom row represents Block 2, both for upright faces. $* P < 0.05$, $** P < 0.01$, $*** P < 0.001$ (one tailed). The color bar represents the magnitude of neural activation, with 0 µV as the minimum, and the participants' individual largest magnitude response as the maximum. Hotter colors represent stronger activation.

from leading theoretical models, such as dominance or shyness (Oosterhof and Todorov, 2008; Collova *et al.*, 2019).

An important open question is which facial cues are driving the implicit neural trustworthiness discrimination response that we observed. It is likely that the trustworthiness dimension represents a combination of values across many different facial attributes (e.g. spacing of facial features, age and masculinity: see Vernon *et al.*, 2014). However, it is important to note that the faces used in the task represent faces chosen to maximally vary on trustworthiness, rather than other key facial dimensions, such as attractiveness or dominance (Todorov et al., 2015). Stimuli were also tightly controlled to minimize other potential confounding factors that may affect facial trustworthiness perception, including face identity, emotion, gender and ethnicity, as well as low-level confounds, such as color, contrast and luminance. Given that FPVS measures whether the brain is responsive to any visual differences between the images presented, this approach was particularly important in the current study. However, the kinds of faces that people see in everyday life are naturally far more varied, and face perception research is increasingly incorporating more naturalistic face images (Jenkins *et al.*, 2011; Sutherland *et al.*, 2013; Lischke *et al.*, 2018). One downside of using tightly controlled stimuli is that it minimizes facial cues that people would typically use to judge trustworthiness (e.g. gender: Sutherland *et al.*, 2015). Future research could use the FVPS technique as a way to further understand trustworthiness perception, by systematically varying facial cues in order to understand their relative importance or by comparing controlled and naturalistic face images (Coll *et al.*, 2019).

## Conclusions

In conclusion, we demonstrated implicit trustworthiness perception using FVPS, advancing an important theoretical debate in the field. Moreover, our study provides an important methodological contribution, as we show that the neural trustworthiness discrimination response is reliable and shows variability at the individual level, opening the door to studies on individual differences and with special populations. As the field moves away from focusing on explicit perceptions (Abir *et al.*, 2018), our study helps to provide new ways of investigating implicit trustworthiness perception and potentially other high-level social face attributes. Taken together, our results demonstrate the robust applicability of FPVS for investigating higher level social attributes, corroborating the growing use of this technique to inform our understanding of face processing, a critical aspect of human social cognition.

## Supplementary data

Supplementary data mentioned in the text are available to subscribers in *SOCAFN* online.

## Acknowledgements

The authors would like to thank Bruno Rossion, Joan Liu-Shuang, Talia Retter and Amy Dawel for their advice and helpful discussions, and Alex Todorov for providing the face stimuli.

## Conflict of Interest

There are no conflicts of interest to report.

## Funding

This research was supported by an RTP scholarship from the University of Western Australia to D.S., an Australian Research Council (ARC) Discovery Early Career Research Award to C.S. [DE190101043], the ARC Centre of Excellence in Cognition and its Disorders [CE110001021], ARC Discovery Awards to G.R., C.S. and R.P. [DP170104602] and R.P. [DP140101743] and by ENS Paris-Saclay to S.P.

## References

Abir, Y., Sklar, A.Y., Dotsch, R., Todorov, A., Hassin, R.R. (2018). The determinants of consciousness of human faces. *Nature Human Behaviour*, **2**(3), 194. doi: 10.1038/s41562-017-0266-3.

Beck, A.A., Rossion, B., Samson, D. (2017). An objective neural signature of rapid perspective taking. *Social Cognitive and Affective Neuroscience*, **13**(1), 72–9. https://doi.org/10.1093/scan/nsx135.

Coll, M.-P., Murphy, J., Catmur, C., Bird, G., Brewer, R. (2019). The importance of stimulus variability when studying face processing using fast periodic visual stimulation: a novel 'mixed-emotions' paradigm. *Cortex*, **117**, 182–95. https://doi.org/10.1016/j.cortex.2019.03.006.

Collova, J.R., Sutherland, C.A.M., Rhodes, G. (2019). Testing the functional basis of first impressions: dimensions for children's faces are not the same as for adults' faces. *Journal of Personality and Social Psychology*, **117**(5), 900–24. http://dx.doi.org/10.1037/pspa0000167.

DeGutis, J., Wilmer, J., Mercado, R.J., Cohan, S. (2013). Using regression to measure holistic face processing reveals a strong link with face recognition ability. *Cognition*, **126**(1), 87–100. https://doi.org/10.1016/j.cognition.2012.09.004.

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), 87. https://doi.org/10.1186/1471-2202-15-87.

Dzhelyova, M., Perrett, D.I., Jentzsch, I. (2012). Temporal dynamics of trustworthiness perception. *Brain Research*, **1435**, 81–90. https://doi.org/10.1016/j.brainres.2011.11.043.

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*, **27**(8), 4106–23. https://doi.org/10.1093/cercor/bhw223.

Engell, A.D., Haxby, J.V., Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, **19**(9), 1508–19. https://doi.org/10.1162/jocn.2007.19.9.1508.

Ewing, L., Sutherland, C.A., Willis, M.L. (2019). Children show adult-like facial appearance biases when trusting others. *Developmental Psychology*, **55**(8), 1694–701. http://dx.doi.org/10.1037/dev0000747.

Gollwitzer, M., Rothmund, T., Süssenbach, P. (2013). The sensitivity to mean intentions (SeMI) model: basic assumptions, recent findings, and potential avenues for future research. *Social and Personality Psychology Compass*, **7**(7), 415–26. https://doi.org/10.1111/spc3.12041.

Gwinn, O.S., Matera, C.N., O'Neil, S.F., Webster, M.A. (2018). Asymmetric neural responses for facial expressions and anti-expressions. *Neuropsychologia*, **119**, 405–16. https://doi.org/10.1016/j.neuropsychologia.2018.09.001.

Hancock, K.J., Rhodes, G. (2008). Contact, configural coding and the other-race effect in face recognition. *British Journal of Psychology*, **99**(1), 45–56. doi: 10.1348/000712607X199981.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, **4**(6), 223–33. https://doi.org/10.1016/S1364-6613(00)01482-0.

Hehman, E., Sutherland, C.A., Flake, J.K., Slepian, M.L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, **113**(4), 513. http://dx.doi.org/10.1037/pspa0000090.

Hunt, R.R. (1995). The subtlety of distinctiveness: what von Restorff really did. *Psychonomic Bulletin & Review*, **2**(1), 105–12. https://doi.org/10.3758/BF03214414.

Jenkins, R., White, D., Van Montfort, X., Burton, A.M. (2011). Variability in photos of the same face. *Cognition*, **121**(3), 313–23. https://doi.org/10.1016/j.cognition.2011.08.001.

Klapper, A., Dotsch, R., van Rooij, I., Wigboldus, D.H. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? *Journal of Personality and Social Psychology*, **111**(5), 655–64. http://dx.doi.org/10.1037/pspa0000090.

Krumhuber, E., Manstead, A.S., Cosker, D., Marshall, D., Rosin, P.L., Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, **7**(4), 730–5. http://dx.doi.org/10.1037/1528-3542.7.4.730.

Lischke, A., Junge, M., Hamm, A.O., Weymar, M. (2018). Enhanced processing of untrustworthiness in natural faces with neutral expressions. *Emotion*, **18**(2), 181–9. http://dx.doi.org/10.1037/emo0000318.

Liu-Shuang, J., Norcia, A.M., Rossion, B. (2014). An objective index of individual face discrimination in the right occipito-temporal cortex by means of fast periodic oddball stimulation. *Neuropsychologia*, **52**, 57–72. https://doi.org/10.1016/j.neuropsychologia.2013.10.022.

Luo, Q., Rossion, B., Dzhelyova, M. (2019). A robust implicit measure of facial attractiveness discrimination. *Social Cognitive and Affective Neuroscience*, **14**(7), 747–6. https://doi.org/10.1093/scan/nsz043.

Mattavelli, G., Andrews, T.J., Asghar, A.U., Towler, J.R., Young, A.W. (2012). Response of face-selective brain regions to trustworthiness and gender of faces. *Neuropsychologia*, **50**(9), 2205–11. https://doi.org/10.1016/j.neuropsychologia.2012.05.024.

McCarthy, G., Wood, C.C. (1985). Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, **62**(3), 203–8. https://doi.org/10.1016/0168-5597(85)90015-2.

Mondloch, C.J., Gerada, A., Proietti, V., Nelson, N.L. (2019). The influence of subtle facial expressions on children's first impressions of trustworthiness and dominance is not adult-like. *Journal of Experimental Child Psychology*, **180**, 19–38. https://doi.org/10.1016/j.jecp.2018.12.002.

Olivola, C.Y., Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, **46**(2), 315–24. https://doi.org/10.1016/j.jesp.2009.12.002.

Olivola, C.Y., Funk, F., Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, **18**(11), 566–70. https://doi.org/10.1016/j.tics.2014.09.007.

Oosterhof, N.N., Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, **105**(32), 11087–92. https://doi.org/10.1073/pnas.0805664105.

Regan, D. (1966). Some characteristics of average steady-state and transient responses evoked by modulated light. *Clinical Neurophysiology*, **20**(3), 238–48. https://doi.org/10.1016/0013-4694(66)90088-5.

Retter, T.L., Rossion, B. (2016). 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., Brake, S., Atkinson, A.P. (1993). What's lost in inverted faces? *Cognition*, **47**(1), 25–57. https://doi.org/10.1016/0010-0277(93)90061-Y.

Rossion, B. (2014). Understanding individual face discrimination by means of fast periodic visual stimulation. *Experimental Brain Research*, **232**(6), 1599–621. https://doi.org/10.1007/s00221-014-3934-9.

Rossion, B., Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews*, **1**(1), 63–75. https://doi.org/10.1177/1534582302001001004.

Rossion, B., Prieto, E.A., Boremanse, A., Kuefner, D., Van Belle, G. (2012). A steady-state visual evoked potential approach to individual face perception: effect of inversion, contrast-reversal and temporal dynamics. *NeuroImage*, **63**(3), 1585–600. https://doi.org/10.1016/j.neuroimage.2012.08.033.

Santos, I., Young, A. (2005). Exploring the perception of social characteristics in faces using the isolation effect. *Visual Cognition*, **12**(1), 213–47. https://doi.org/10.1080/13506280444000102.

Schmitt, M., Baumert, A., Gollwitzer, M., Maes, J. (2010). The justice sensitivity inventory: factorial validity, location in the personality facet space, demographic pattern, and normative data. *Social Justice Research*, **23**(2–3), 211–38. https://doi.org/10.1007/s11211-010-0115-2.

Schweinberger, S.R., Neumann, M.F. (2016). Repetition effects in human ERPs to faces. *Cortex*, **80**, 141–53. https://doi.org/10.1016/j.cortex.2015.11.001.

Stothart, G., Quadflieg, S., Milton, A. (2017). A fast and implicit measure of semantic categorisation using steady state visual evoked potentials. *Neuropsychologia*, **102**, 11–8. https://doi.org/10.1016/j.neuropsychologia.2017.05.025.

Sutherland, C.A., Oldmeadow, J.A., Santos, I.M., Towler, J., Burt, D.M., Young, A.W. (2013). Social inferences from faces: ambient images generate a three-dimensional model. *Cognition*, **127**(1), 105–18. https://doi.org/10.1016/j.cognition.2012.12.001.

Sutherland, C.A., Young, A.W., Mootz, C.A., Oldmeadow, J.A. (2015). Face gender and stereotypicality influence facial trait evaluation: counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology*, **106**(2), 186–208. https://doi.org/10.1111/bjop.12085.

Sutherland, C.A., Rhodes, G., Williams, N., *et al.* (2019). Appearance-based trust processing in schizophrenia. *British Journal of Clinical Psychology*. doi: https://doi.org/10.1111/bjc.12234.

Taylor, S.E., Fiske, S.T., Etcoff, N.L., Ruderman, A.J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, **36**(7), 778–93. http://dx.doi.org/10.1037/0022-3514.36.7.778.

Todorov, A., Said, C.P., Oosterhof, N.N., Engell, A.D. (2011). Task-invariant brain responses to the social value of faces. *Journal of Cognitive Neuroscience*, **23**(10), 2766–81. https://doi.org/10.1162/jocn.2011.21616.

Todorov, A., Dotsch, R., Porter, J.M., Oosterhof, N.N., Falvello, V.B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, **13**(4), 724–38. https://psycnet.apa.org/doi/10.1037/a0032335.

Todorov, A., Olivola, C.Y., Dotsch, R., Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual review of psychology*. **66**, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831.

Van't Wout, M., Sanfey, A.G. (2008). Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. *Cognition*, **108**(3) 796, 796–803. https://doi.org/10.1016/j.cognition.2008.07.002.

Vernon, R.J., Sutherland, C.A., Young, A.W., Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, **111**(32), E3353–E3361. https://doi.org/10.1073/pnas.1409860111.

Von Restorff, H. (1933). Über die wirkung von bereichsbildungen im spurenfeld. *Psychologische Forschung*, **18**(1), 299–342. https://doi.org/10.1007/BF02409636.

Walker, M., Vetter, T. (2016). Changing the personality of a face: perceived big two and big five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, **110**(4), 609–24. http://dx.doi.org/10.1037/pspp0000064.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G.O., Gosselin, F., Tanaka, J.W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior Research Methods*, **42**(3), 671–84. https://doi.org/10.3758/BRM.42.3.671.

Willis, J., Todorov, A. (2006). First impressions: making up your mind after a 100-ms exposure to a face. *Psychological Science*, **17**(7), 592–8. https://doi.org/10.1111%2Fj.1467-9280.2006.01750.x.

Winston, J.S., Strange, B.A., O'Doherty, J., Dolan, R.J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, **5**(3), 277–83. http://dx.doi.org/10.1038/nn816.

Yang, D., Qi, S., Ding, C., Song, Y. (2011). An ERP study on the time course of facial trustworthiness appraisal. *Neuroscience Letters*, **496**(3), 147–51. https://doi.org/10.1016/j.neulet.2011.03.066.

Yovel, G., Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, **15**(24), 2256–62. https://doi.org/10.1016/j.cub.2005.10.072.

Zebrowitz, L.A. (2017). First impressions from faces. *Current Directions in Psychological Science*, **26**(3), 237–42. https://doi.org/10.1177/0963721416683996.

Zhu, M., Alonso-Prieto, E., Handy, T., Barton, J. (2016). The brain frequency tuning function for facial emotion discrimination: an ssVEP study. *Journal of Vision*, **16**(6), 12–2. doi: 10.1167/16.6.12.