# Trustworthiness perception is mandatory: Task instructions do not modulate fast periodic visual stimulation trustworthiness responses

**Derek C. Swe** School of Psychological Science, The University of Western Australia, Perth, Australia

**Romina Palermo** School of Psychological Science, The University of Western Australia, Perth, Australia

**O. Scott Gwinn** College of Education, Psychology, and Social Work, Flinders University, Adelaide, Australia

**Jason Bell** School of Psychological Science, The University of Western Australia, Perth, Australia

**Anju Nakanishi** School of Psychological Science, The University of Western Australia, Perth, Australia

**Jemma Collova** School of Psychological Science, The University of Western Australia, Perth, Australia

**Clare A. M. Sutherland** School of Psychological Science, The University of Western Australia, Perth, Australia
School of Psychology, University of Aberdeen, King's College, Aberdeen, Scotland

**Although it is often assumed that humans spontaneously respond to the trustworthiness of others' faces, it is still unclear whether responses to facial trust are mandatory or can be modulated by instructions. Considerable scientific interest lies in understanding whether trust processing is mandatory, given the societal consequences of biased trusting behavior. We tested whether neural responses indexing trustworthiness discrimination depended on whether the task involved focusing on facial trustworthiness or not, using a fast periodic visual stimulation electroencephalography oddball paradigm with a neural marker of trustworthiness discrimination at 1 Hz. Participants judged faces on size without any reference to trust, explicitly formed impressions of facial trust, or were given a financial lending context that primed trust, without explicit trust judgement instructions. Significant trustworthiness discrimination responses at 1 Hz were found in all three conditions, demonstrating the robust nature of trustworthiness discrimination at the neural level. Moreover, no effect of task instruction was observed, with Bayesian analyses providing moderate to decisive evidence that task instruction did not affect trustworthiness discrimination. Our finding that visual trustworthiness discrimination is mandatory points to the remarkable spontaneity of trustworthiness processing, providing clues regarding why these often unreliable impressions are ubiquitous.**

## Introduction

Face-based trust judgements can have widespread social implications, affecting economic decisions, corporate success, and even legal outcomes, (Jaeger, Todorov, Evans, & van Beest, 2020; Linke, Saribay, & Kleisner, 2016; Olivola, Funk, & Todorov, 2014; Olivola & Todorov, 2010). There are links between trustworthiness impressions and the likelihood of

Citation: Swe, D. C., Palermo, R., Gwinn, O. S., Bell, J., Nakanishi, A., Collova, J., & Sutherland, C. A. M. (2022). Trustworthiness perception is mandatory: Task instructions do not modulate fast periodic visual stimulation trustworthiness responses. *Journal of Vision*, *22*(11):17, 1–19, https://doi.org/10.1167/jov.22.11.17.

https://doi.org/10.1167/jov.22.11.17 Received October 17, 2021; published October 31, 2022 ISSN 1534-7362 

in-group acceptance and inclusion (Tracy, Wilson, Slepian, & Young, 2020), as well as a relationship between facial trustworthiness and position in corporate hierarchy (Linke et al., 2016). These studies demonstrate how many different facets of life can be impacted by trustworthiness impressions and just how strong these impacts can be. As a result, there has been growing scientific interest in understanding how facial trustworthiness is perceived, and to what extent these (often biased) judgements can be mitigated or changed (Brambilla, Biella, & Freeman, 2018; Jaeger et al., 2020; Sutherland, Burton, Wilmer, Blokland, Germine, Palermo, Collova, & Rhodes, 2020).

Theories of trustworthiness perception propose that trustworthiness judgements are largely made automatically (Dzhelyova, Perrett, & Jentzsch, 2012; Eggleston, Flavell, Tipper, Cook, & Over, 2021; Marzi, Righi, Ottonello, Cincotta, & Viggiano, 2014; Todorov, Said, Engell, & Oosterhof, 2008; Zebrowitz, 2017). One reason that trustworthiness may be judged automatically is because these judgments have been argued to be functional (Collova, Sutherland, & Rhodes, 2019; Oosterhof & Todorov, 2008; Willis & Todorov, 2006). That is, in an evolutionary sense, the ability to visually discriminate between trustworthy and untrustworthy individuals could have played an important role in adaptive threat detection, either directly, or, more likely, given the low accuracy of these judgements, as a by-product of other face perception processes, such as overgeneralization from emotion recognition (Oosterhof & Todorov, 2008; Zebrowitz, 2004). More generally, if these facial biases are automatic, they may help minimize cognitive load, similarly to other types of stereotyping (Siddique, Jeffery, Palermo Collova, Sutherland, 2022).

For a process to be considered automatic, it should fulfil some or all of the four main requirements: it should be capacity-free (made readily), non-conscious, rapid, and mandatory (occurring regardless of intention) (Palermo & Rhodes, 2007). It is still unclear exactly how these facets of automaticity are related. This ambiguity has led to past theoretical discussions about what it means for a process to be automatic (Moors & De Houwer, 2012), and how automaticity may apply to face perception (Palermo & Rhodes, 2007). Both Moors and De Houwer (2012) and Palermo and Rhodes (2007) suggest that processes can be automatic in many ways, and that studies should decide on the specific aspect of automaticity they are investigating. Previous landmark studies have already shown that some aspects of automaticity apply for trustworthiness processing; judgements of trustworthiness from faces are made readily (Oosterhof & Todorov, 2008; Sutherland, Liu, Zhang, Chu, Oldmeadow, & Young, 2018), unconsciously (Freeman, Stolier, Ingbretsen, & Hehman, 2014; Stewart, Ajina, Getov, Bahrami, Todorov, & Rees, 2012) and quickly, within milliseconds (Dzhelyova et al., 2012; Willis & Todorov, 2006). The findings from these studies have informed current theory, providing key evidence that trustworthiness may be processed automatically. However, there is one important aspect of automaticity that has been less directly examined with reference to trustworthiness. It is still unclear whether these judgements are mandatory. That is, do trustworthiness judgements occur regardless of people's intention or motivation to judge trust? Thus the current study will answer this important question by investigating the mandatory nature of trustworthiness processing.

Converging behavioral, electrophysiological, and neuroimaging research has indicated that trustworthiness perception can occur implicitly, that is without needing explicit impression formation instructions (Klapper, Dotsch, van Rooij, & Wigboldus, 2016; Winston, Strange, O'Doherty, & Dolan, 2002; Swe, Palermo, Gwinn, Rhodes, Neumann, Payart, & Sutherland, 2020; Verosky, Zoner, Marble, Sammon, & Babarinsa, 2020), although not all studies do find evidence for implicit trustworthiness perception (Santos & Young, 2005). The fact that trustworthiness is perceived implicitly from faces is suggestive of mandatory processing of trust. However, without a direct comparison between implicit and explicit responses, it is unclear whether trust responses can be modulated by task or resulting intentions, and thus whether trust perception from faces is truly mandatory. Interestingly, a recent study using explicit judgements of trustworthiness demonstrated that the influence of facial stereotypes (e.g., facial trustworthiness) cannot be mitigated even when people are educated about biasing effects of facial stereotyping (Jaeger et al., 2020), suggesting that trustworthiness processing may be a mandatory process that can persistently affect social decision-making. However, these studies, although suggestive of mandatory processing, have not directly compared an explicit trustworthiness judgement task alongside an implicit task. Therefore it is difficult to draw conclusions about whether these impressions are truly mandatory because it is not clear whether participants would have formed the same judgements regardless of whether participants were focusing on trustworthiness or not.

Moreover, other behavioral evidence suggests that trustworthiness impressions can be influenced by the context to a large extent, suggesting that trust judgements are not necessarily mandatory. In an important new line of work, Brambilla and colleagues (2018) showed that the threatening or nonthreatening nature of a visual scene around a face can alter reaction times when participants are asked to judge trustworthiness from faces, suggesting that (visual) context can have an implicit influence on trustworthiness processing. Similarly, another recent

study found that contextual auditory cues can also modulate trustworthiness impressions (Brambilla, Masi, Mattavelli, & Biella, 2021).

More direct evidence comes from neuroimaging studies, which have found that task instructions modulated neural responses to trustworthiness, although not consistently so (Marzi et al., 2014; Winston et al., 2002). For example, although the bilateral amygdala and right insula responded to untrustworthy faces regardless of task instruction, the right superior temporal sulcus showed stronger responses when explicit trustworthiness judgement instructions were given (Winston et al., 2002). Similarly, certain event-related potentials have been shown to be enhanced during trustworthiness discrimination, but not during political decision making (Marzi et al., 2014). Although these studies have provided the most direct test of mandatory processing to date, they are restricted in their ability to provide strong evidence of mandatory processing. Both age (Winston et al., 2002) and political judgements (Marzi et al., 2014) potentially cue trustworthiness (for example, older faces look more trustworthy: Sutherland, Oldmeadow, Santos, Towler, Burt, & Young, 2013) and thus do not necessarily allow for implicit processing. Nevertheless, these neuroimaging and behavioral findings are an important first step to suggest that facial trustworthiness impressions can (in some circumstances) be modified, and that internal goals and motivations, affected by contextual cues or task instructions, may influence these impressions.

In summary, only a few studies have examined the influence of task instruction on trustworthiness processing, and the findings have been inconsistent across different paradigms. Moreover, there has been no direct investigation into whether trustworthiness processing is mandatory in an experiment with both explicit and implicit task instructions. It is crucial to determine whether facial trustworthiness processing is mandatory or malleable because understanding to what extent responses can be shifted is an important aspect of automaticity that has yet to be thoroughly understood for face trustworthiness processing. Furthermore, it is critical to understand to what extent we can shift trustworthiness processing given the important social implications of these judgements (Jaeger et al., 2019; Sutherland et al., 2020; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015).

We set out to investigate whether trustworthiness processing is mandatory by taking advantage of the fast periodic visual stimulation (FPVS) technique, an advance in the field of electroencephalography (EEG). To investigate this question, we used the FPVS oddball technique, taken from two recent studies which have shown that trustworthiness can be processed implicitly (Swe et al., 2020; Verosky et al., 2020). This FPVS oddball paradigm (Rossion, 2014) involves sequentially presenting faces at a predetermined, base frequency (e.g., six faces per second, resulting in a base frequency of 6 Hz). Within this sequence, an attribute of interest (here, trustworthiness) changes at a different, oddball frequency (e.g., every sixth face differs in apparent trustworthiness compared to the other five, resulting in an oddball frequency of 1 Hz). In the paradigm used by Swe et al. (2020), participants were tasked only to attend to a fixation cross and respond when it changed; there were no explicit impression formation instructions. Swe et al. (2020) found a significant oddball response in the visual cortex corresponding with face trustworthiness changes (i.e., at the 1 Hz oddball frequency). This result suggests that facial trustworthiness can be processed implicitly, without needing instructions to judge trustworthiness. Using a similar FPVS oddball face individuation paradigm, Verosky et al. (2020) also found evidence of implicit trustworthiness processing. Together these results suggest that the FPVS oddball response can be used as an objective and reliable neural marker of trustworthiness processing. However, because neither experiment had a task-directed condition, it is still unclear whether these responses can be changed by asking people to judge trustworthiness explicitly. For face *identity* perception, a recent FPVS study has shown that when participants were given a face-related task during the presentation of faces, a stronger face identity discrimination response was found compared to when a non-face-related task was given (Yan, Liu-Shuang, & Rossion, 2019). A similar question could be posed for trustworthiness processing.

Here, we adapted the FPVS trust paradigm to investigate whether the neural trustworthiness discrimination response changes when given trust-relevant instructions. FPVS is ideal for addressing this question because responses to trustworthiness can be measured in the absence of instructions and because the technique itself has a high level of objectivity: the predictions are a priori, the frequencies are predetermined, avoiding any potential issues with differences in electrophysiological components not under the control of the experimenter, and results are clear-cut (one either finds a significant oddball response, plus associated harmonics, or not).

If trustworthiness processing is mandatory, we would predict that task instructions to judge the faces' trustworthiness does not modulate the neural trustworthiness discrimination. However, if trustworthiness processing is instead malleable, we would predict that the neural trustworthiness discrimination response is less strong when the attended attribute is irrelevant to trustworthiness (i.e., when participants are tasked to judge the size of the faces) compared with when the attributes are relevant (i.e., when participants are tasked to judge the trustworthiness of the faces). To test these questions, we contrasted neural trustworthiness discrimination

responses to faces when participants were asked to make judgements about the size of faces (which varied orthogonally to trustworthiness) versus their responses when asked to explicitly consider whether the faces were trustworthy or untrustworthy.

Additionally, we also tested a third condition where trustworthiness was primed rather than explicitly stated. In this condition, we asked participants to decide whether or not to lend money to the faces shown. Previous research has shown that the context of lending money to strangers makes trustworthiness a salient dimension for participants to consider (Rezlescu, Duchaine, Olivola, & Chater, 2012; Van't Wout & Sanfey, 2008). In this way, we can test whether individuals show stronger neural trustworthiness discrimination responses when internally motivated, as compared to when explicitly instructed to focus on trustworthiness, or when trust is not mentioned at all. Looking at trustworthiness in this way also presents greater ecological validity, as real-life contexts do not often involve being asked to explicitly judge trustworthiness, even if trustworthiness is clearly important (such as when choosing to lend money).

All participants first completed the implicit size condition so that our comparison of task instructions benefitted from a within-participant design. In this initial block, we also aimed to replicate the recent finding that an FPVS signal can be found in response to faces changing in trustworthiness in the absence of explicit trust judgement instructions (Swe et al., 2020; Verosky et al., 2020).

Crucially, the face sequences shown in each condition were identical. The critical aspect that changed between conditions was the instructions regarding which feature to attend to: face image size in the size condition, face trustworthiness in the trust condition, and desire to lend money in the economic context condition. Thus any differences in neural response between conditions cannot be attributable to differences between the face stimuli, but rather to task instructions. If neural responses are similar across conditions, it would indicate that the neural trustworthiness discrimination response is likely mandatory. However, if the strength of the neural response is higher in the conditions in which trust is relevant (explicit trust and contextually-relevant trust conditions) compared to the trust irrelevant (implicit trust) condition, this pattern instead suggests that the neural trustworthiness discrimination response can be modulated by task instructions. In this case, the results would also provide evidence of a link between the internal motivations or goals of the individual and visual processing of trustworthiness. Finally, it is also possible that the trust-relevant economic context would increase sensitivity to trustworthiness to a lesser extent than explicit judgements of trustworthiness did, suggesting a graded responsiveness. To test these alternative predictions, we used a Bayesian approach, because we were interested in the strength of evidence for or against any difference among the three conditions.

# Materials and methods

The methods and analyses were pre-registered on OSF (https://osf.io/us7bf) and were followed during this study. A few deviations were made. Although the majority of participants were randomly assigned to each condition, as per the preregistration, because of a highly biased number of males in one condition (noticed after 66 participants were tested), the last 20 participants were pseudo-randomly assigned to a condition, with males being assigned to the two conditions (trust and economic context) which had less males. Additionally, the pre-registered report mistakenly suggested that participants who blinked more than 0.2 times per second would be removed. Instead, in the present study, participants who blinked more than 0.2 times per second had blink corrections applied to their data, as per Swe et al. (2020). These changes were made before carrying out any statistical testing.

## Participants

The final sample consisted of 86 participants (37 males, ages ranging from 18 to 60, M = 23.80 years, SD = 8.94 years). Sample size was based on multiple power analyses (conducted in R, version 3.6.1). Based on a previous face perception FPVS study (Beck, Rossion, & Samson, 2017), a power analysis conducted in G*power found that 28 participants were needed to find an effect size of .29 with a power of .8 at the standard .05 alpha error probability. Additionally, a power analysis was run based on a recent FPVS study looking at task modulation on individual face discrimination (Yan et al., 2019), which found that 12 participants would be needed to find an effect size of 0.66 with a power of 0.99 at the conservative 0.01 alpha probability for a repeated-measures analysis of variance (ANOVA). Therefore we aimed to test 30 participants per group for a final sample size of 90. Because four participants were excluded from the data analysis due to missing data, the final sample was 86, and each condition met our minimum requirement of 28 participants. The size condition consisted of 29 participants (15 males, ages ranging from 18 to 56, $M = 21.90$ years, $SD = 7.64$ years), the trust condition consisted of 29 participants (12 males, ages ranging from 18 to 56, $M = 25.41$ years, $SD = 9.88$ years), and the economic context condition consisted of 28 participants (10 males, ages ranging from 18 to 60, $M = 24.11$ years, $SD = 9.10$ years). Our sample size for each condition is also

comparable to that of Swe et al. (2020), who had 31 participants.

Participants were students at the University of Western Australia ($N = 66$) or recruited from the wider community ($N = 20$). Only Caucasian participants were tested to control for potential other-race effects (Hancock & Rhodes, 2008; Meissner & Brigham, 2001), because Caucasian (computer-generated) face stimuli were used. The study was approved by the University of Western Australia human ethics committee.

## Stimuli

Twenty pairs of FaceGen faces, each consisting of a trustworthy and an untrustworthy version of the same original face identity, were taken from Swe et al. (2020) (originally from Todorov, Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, 2013). These images were originally modelled with FaceGen software, where each face was represented as a point in a face space with 100 dimensions (50 shape and 50 reflectance dimensions). Social dimensions such as trustworthiness were represented and modelled using linear combinations of basic FaceGen dimensions based on trait judgements (Todorov et al., 2013). For example, a dimensional value of 0 on trustworthiness would indicate that the face is neutral trustworthy (neither trustworthy nor untrustworthy), whereas positive and negative values would indicate that the face is more or less trustworthy, respectively. Critically, faces that were morphed to represent trustworthiness varied maximally on the trustworthiness dimension (compared to any other dimension such as dominance, likability, or attractiveness). Trustworthy faces in the current study had a level of 1 SD on the trustworthiness dimension, whereas untrustworthy faces had a level of −3 SD. This asymmetry was a precaution, because of the concern that increasing trustworthiness further made the male faces start to look androgynous or female.

Faces were adjusted to control for low-level influences (luminance, contrast, and grayscale controlled) following the same procedures as Swe et al. (2020). Face image size was jittered in 2% steps in each sequence. This jitter was added to reduce the potential impact of low-level properties of the images contributing to any observed effects (Dzhelyova & Rossion, 2014) and also allowed for judgments to be made of the faces that are unrelated to trustworthiness. For half of the participants, 12 of the 20 face sequences had sizes ranging from 70% to 110%, and eight sequences ranging from 90% to 130%, and vice versa for the other half of participants. Participants were randomly assigned to either the majority smaller or majority larger group.

## Procedure

An FPVS oddball paradigm was used (Liu-Shuang, Norcia, & Rossion, 2014) following the design from Swe et al. (2020); see Figure 1. Faces were shown at a base rate of six faces per second (6 Hz), with oddball faces shown at every sixth face resulting in an oddball frequency of 1 Hz, for 40 seconds of stimulation. All 20 face identities were shown equally often as base and oddball faces across the different sequences to avoid trustworthiness being confounded with identity. Each sequence lasted 40 seconds, 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 sequence).

The task consisted of two blocks. In the first block, all participants completed the size discrimination task. Before the task began, participants were given instructions to verbally respond with "small" or "large" at the end of each face sequence, depending on whether they thought the faces in the sequence were small or large overall. Because participants judged a feature unrelated to trustworthiness, the oddball response to the faces in this condition reflects implicit trustworthiness discrimination. This task has been adapted from the FPVS paradigm used in Swe et al. (2020), and we expected to observe an implicit response to trustworthiness, in replication.

In the second block, participants completed one of three conditions (Figure 1). One third of participants repeated the "size" (implicit trust) condition, which was identical to the first block (i.e., participants were asked to judge the size of the faces). This design allowed us to directly compare the blocks to check for any order or learning effects, and to measure test-retest reliability. In the "explicit trust" condition, participants were explicitly instructed to judge the trustworthiness of the faces. Participants were asked to verbally respond with "trustworthy" or "untrustworthy" at the end of each face sequence depending on whether they thought the faces, overall, belonged to the trustworthy category or untrustworthy category. As the task was an oddball paradigm, half the sequences included majority trustworthy faces (i.e., in untrustworthy oddball sequences, there are five trustworthy faces and one untrustworthy face), and vice versa. Therefore, the oddball response to the faces in this condition should reflect explicit trustworthiness discrimination, with the possibility of the strength of the response being increased (as compared to implicit size judgements) by explicit attention to trustworthiness.

Finally, we included an "economic context condition," because context has previously been shown to impact face perception (Barrett, Mesquita, & Gendron, 2011; Brambilla et al., 2018; Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011). In this economic context condition, participants were

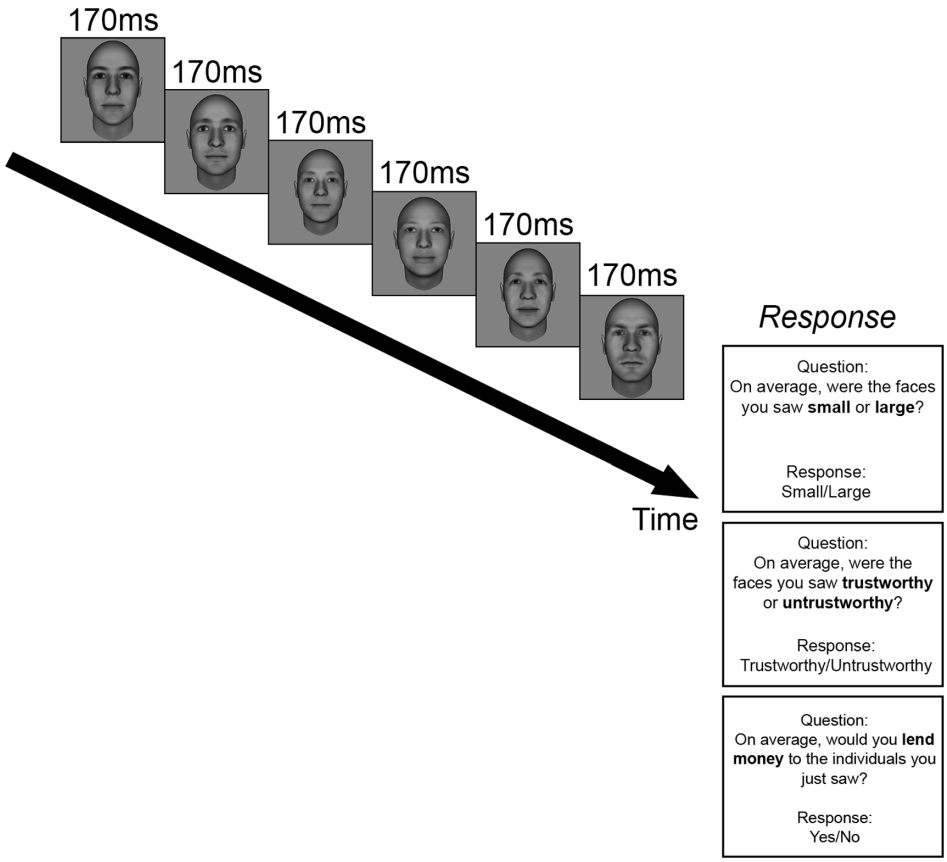


Figure 1. Example untrustworthy oddball FPVS sequence. Each block contained ten 40-second sequences in which faces were presented at a frequency of 6 Hz, with oddball images shown at a frequency of 1 Hz. Thus an untrustworthy oddball sequence consisted of five trustworthy base (T) face images followed by one untrustworthy oddball (UT) (and vice versa for trustworthy oddball sequences). Participants responded at the end of each sequence. 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. Stimuli originally digitally generated by Todorov et al. (2013) and luminance, contrast, and grayscale controlled by Swe et al. (2020).

given $10 (that they could keep) and had to decide whether they would lend that money to the faces shown during each sequence of the task by verbally respond with "yes" or "no" at the end of each face sequence depending on whether they would lend the money. The oddball response to the faces in the economic context condition should again reflect implicit trustworthiness perception (at least so that participants are not asked to explicitly judge trustworthiness), but with the possibility of the strength of the response being increased by a context where trustworthiness cues are more salient.

Using two blocks in this way also allows the data to be analyzed within-subjects (block 1 vs. block 2 for all three conditions), but also between-subjects (block 2 in one condition vs. block 2 in another condition). In addition, half of the sequences involved inverted faces, which were identical to the upright sequences but with the faces rotated 180°. Because inverting a face disrupts normal face processing, while keeping processing of low-level features intact (Rhodes et al., 1993; Rossion & Gauthier, 2002; Yovel & Kanwisher, 2005), using the inverted sequences as a comparison to the upright serves to test whether the neural response was face-selective.

Faces were presented using a square wave function with a 100% duty cycle (Swe et al., 2020). That is, within a sequence, 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 after. The computer refresh rate was set to 60 Hz.

After the FPVS tasks were completed, participants were asked to explicitly rate the trustworthiness of each face used in the FPVS tasks on a scale of 1–9 (1 = not at all trustworthy to 9 = extremely trustworthy) using Qualtrics (Provo, UT, USA). There was no time limit on responding and each face remained visible on the screen until a response was made. The rating task served as a manipulation check. The experiment took

approximately 60 minutes, including 15 minutes to set up the EEG.

## 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, Inc., Natick, MA, USA). Standard FPVS processing procedures were followed (Gwinn, Matera, O'Neil, & Webster, 2018; Retter & Rossion, 2016; Rossion, Prieto, Boremanse, Kuefner, & Van Belle, 2012; Swe et al., 2020) (see Supplementary Figure S1 in the supplementary materials for a graphical representation of the processing steps). The EEG data were initially bandpass-filtered at a high-pass cutoff of 0.1 Hz and a low-pass cutoff 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 to the exact presentation duration (0 to 40 seconds). Seventeen participants were identified who on average blinked more than 0.2 times per second during the 40-second stimulation sequences. This criterion was chosen based on previous FPVS studies (Swe et al., 2020; Gwinn et al., 2018; Retter & Rossion, 2016). For these individuals, blink corrections were applied using an independent component analysis with a square matrix (Retter & Rossion, 2016). Across all participants, 14 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. Only one participant required interpolation of a region of interest (ROI) channel (P8).

Individual channels were re-referenced to the average of all 64 electrodes, and waveforms were averaged across both trustworthy and untrustworthy oddball conditions to improve the signal-to-noise ratio of the recordings. The data were then averaged across sequences, keeping the upright and inverted conditions separate, resulting in two waveforms for each participant for each block. These waveforms were then transformed to the frequency domain using a fast Fourier transform. When comparing responses across participants and conditions, baseline corrections were applied, following Swe et al. (2020). This correction 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 ($x' = x - \text{baseline}$). 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 ($z = (x - \text{baseline})/\text{standard deviation of the baseline}$). In this instance, local maximum and minimum amplitude bins were included in the baseline (Rossion et al., 2012; Srinivasan, Russell, D. P., Edelman, G. M., & Tononi, 1999).

We focused our analysis on a predefined right occipitotemporal ROI, comprising electrodes over scalp regions previously shown to be associated with face processing (electrodes P8, P10, and PO8: Dzhelyova & Rossion, 2014; Retter & Rossion, 2016). This choice was motivated by the recent FPVS findings of Swe et al. (2020) and Verosky et al. (2020), which showed evidence of a neural marker for trustworthiness perception in the right occipitotemporal region, as well as from recent fMRI literature suggesting that occipitotemporal cortex may subserve these higher-order judgements (Mattavelli, Andrews, Asghar, Towler, & Young, 2012; Todorov, Said, Oosterhof, & Engell, 2011). We expected that the visual cortex would be sensitive to facial trustworthiness in all three conditions, and should induce an amplitude spike in neural activity in the right occipitotemporal region of interest, at the oddball frequency and associated harmonics (i.e., a neural trustworthiness discrimination response).

When analyzing each condition, the signal-to-noise ratio spectra were grand-averaged across participants, separately for each of the six conditions (three judgement conditions, upright and inverted). For significance testing, z scores were computed from the amplitude spectra grand-averaged across participants, separately for each condition. When comparing amplitudes between judgement conditions, 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 eighth harmonic, i.e., 8 Hz) were summed separately upright and inverted conditions. We excluded the 6 Hz presentation frequency because this frequency represents the general visual response to the faces (i.e., the presence of any face stimulus).

# Results

## Frequency domain quantification of EEG responses

All statistical analyses were run in JASP (JASP Version 0.14.1, JASP Team, 2018). The trustworthiness neural discrimination response was first quantified using data from the first block of the FPVS task across all participants. Z scores of the fundamental oddball frequency (1 Hz) and its harmonics for the first block are shown in Table S1. Significant responses were observed at the fundamental oddball frequency of 1 Hz in the right occipitotemporal region for upright faces ($z = 2.20$, $p = 0.014$) but not inverted faces ($z = .17$, $p = 0.433$). Significant responses were also found at the harmonics (up to 8 Hz) for the upright faces, as well as the inverted faces (see Supplementary Tables S1 and S2 in the supplementary materials for $z$ scores of each harmonic). Pairwise $t$-tests were performed to estimate the mean and 95% confidence intervals (CI, displayed in square parentheses) of the difference between the upright and inverted conditions. CIs were calculated in JASP. The summed oddball response (sum of the baseline subtracted amplitudes for the fundamental frequency and its significant harmonics) was significantly higher in the upright ($M = .30$, $SD = .32$) compared to the inverted ($M = .07$, $SD = .21$) condition across participants in the first block, $t(85) = 5.89$, $p < .001$, [.15, .30], indicating that low-level visual stimuli differences cannot account for the trustworthiness neural discrimination response. Similarly, the strength of the response (baseline subtracted amplitudes) at the fundamental frequency, examined without harmonics, was also significantly higher in the upright ($M = .01$, $SD = .21$) compared to the inverted ($M = -.0002$, $SD = .15$) condition across participants in the first block, $t(85) = 3.55$, $p < .001$, [.04, .15]. These results are consistent with Swe et al. (2020). Oddball responses and scalp topographies for both upright and inverted conditions in the first block are shown in Figure 2. Scatter plots for oddball responses in the orientation

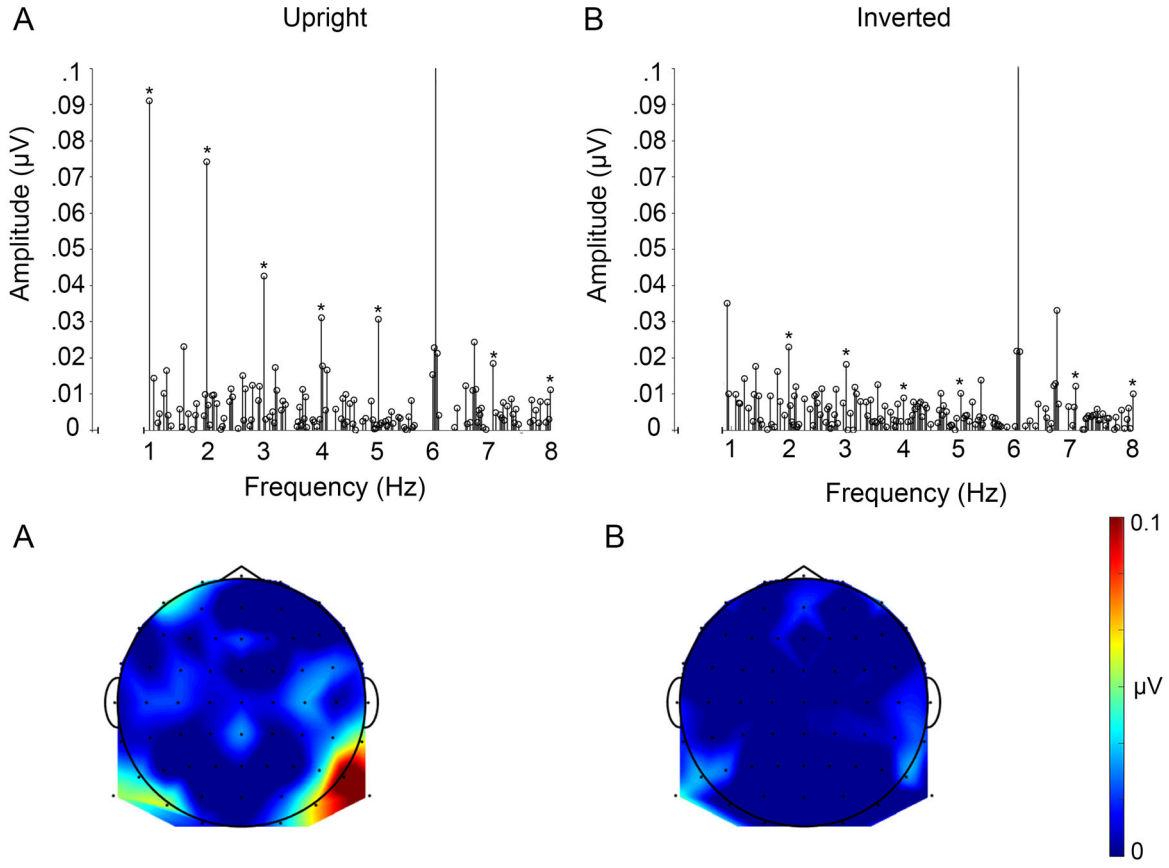


Figure 2. Oddball response amplitude spectra and scalp topographies for the (A) upright and (B) inverted conditions for the first block across all participants ($N = 86$). Top row: baseline subtracted amplitude spectra, collapsed across both trustworthy and untrustworthy oddball face stimuli at the ROI (consisting of electrodes P8, PO8, and P10). Bottom row: scalp topographies for the trustworthiness oddball response (1 Hz), grand averaged across participants. * $p < .05$. All z-value tests report one-tailed $p$ values, which are appropriate here because the signal is only ever meaningful above zero.

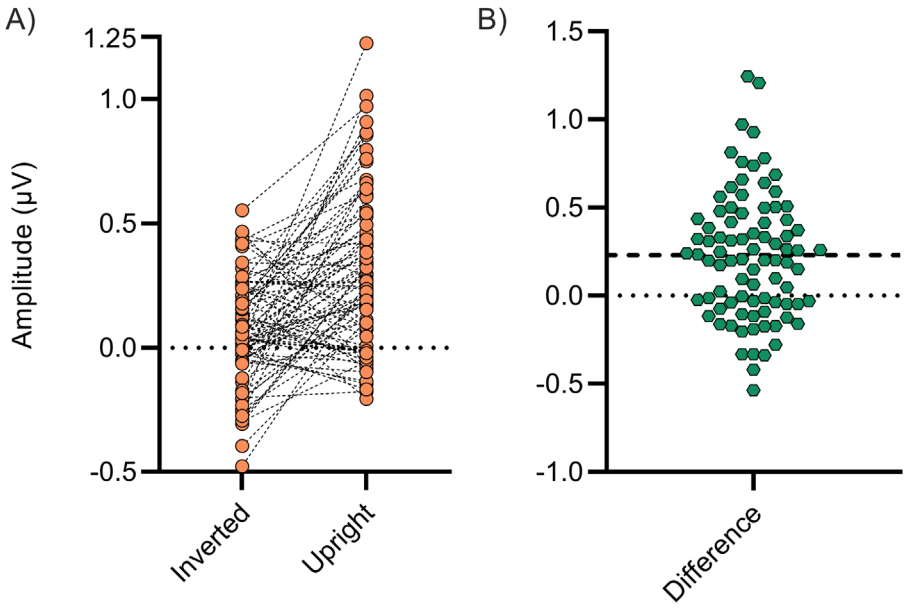


Figure 3. Scatter plots with individual data points of the summed oddball response for each orientation condition in the first block across all participants. Part A shows the upright and inverted conditions, and part B shows the difference (inverted subtracted from upright), ($N = 86$). The *dashed line* represents the mean difference. Overall, the majority of participants ($n = 57$) showed a difference value that was greater than 0.

conditions and their pairwise differences are shown in Figure 3.

## Task modulation on the neural discrimination response

The aim of this study was to determine whether task instruction modulates the neural trustworthiness discrimination response by comparing the strength of the response between conditions. We analyzed this data using Bayesian statistics as the primary analysis and classical statistics as a complementary analysis, consistent with our pre-registration plan. The Bayes' approach focuses on the strength of the evidence for any similarities or differences, and is complemented by the widely used classical statistics approach which focuses on finding potential differences. Assumption checks were conducted in accordance with Goss-Sampson et al. (2020). Tests of normality suggested that assumptions for both the Bayesian and classical statistics were met, as there were no outliers, each block in each condition had adequately normal distributions (skews −0.06 to 0.84, kurtosis −0.87 to 0.09, see Supplementary Figure S2 in the supplementary materials for Q-Q plots), and Levene's test indicated homogeneity of variance ($F = 1.14$, $p = 0.33$). For the Bayesian analyses, the Bayes factor ($BF_{10}$), which shows the strength of evidence in favor of the alternative compared to the null model (Jeffreys, 1961), was reported.

To determine the strength of evidence for or against a difference between conditions, we conducted a Bayesian mixed factor ANOVA. The results from the ANOVA indicated that the data were best represented by a model that included the main factor of orientation only. The Bayes factor ($BF_{10}$) was $3.13e^{13}$, indicating decisive evidence in favor of this model when compared to the null model. These results reiterate that the trustworthiness neural discrimination response is stronger for upright faces compared to inverted faces, indicating a strong face selectivity in the response.

The Bayes factors for block and condition were .13 and .16 respectively, indicating moderate evidence in favor of the null model (i.e., moderate evidence that the neural response was not different between blocks or conditions). The inclusion of Bayes factors for the two-way block*orientation, block*condition, and condition*orientation interactions were 0.05, 0.01, and 0.02 respectively, suggesting strong evidence against the inclusion of these two-way interactions as predictors in a model. The inclusion Bayes factor for the three-way block*condition*orientation interaction was $8.81*e^{-5}$ suggesting decisive evidence against the inclusion of the three-way interaction as a predictor in the model, and that the data are 11,356.24 times more likely to exist under models that exclude the three-way interaction than under models that include this predictor. Critically, these results suggest that there is moderate to decisive evidence that the neural trust response did not differ between task instruction

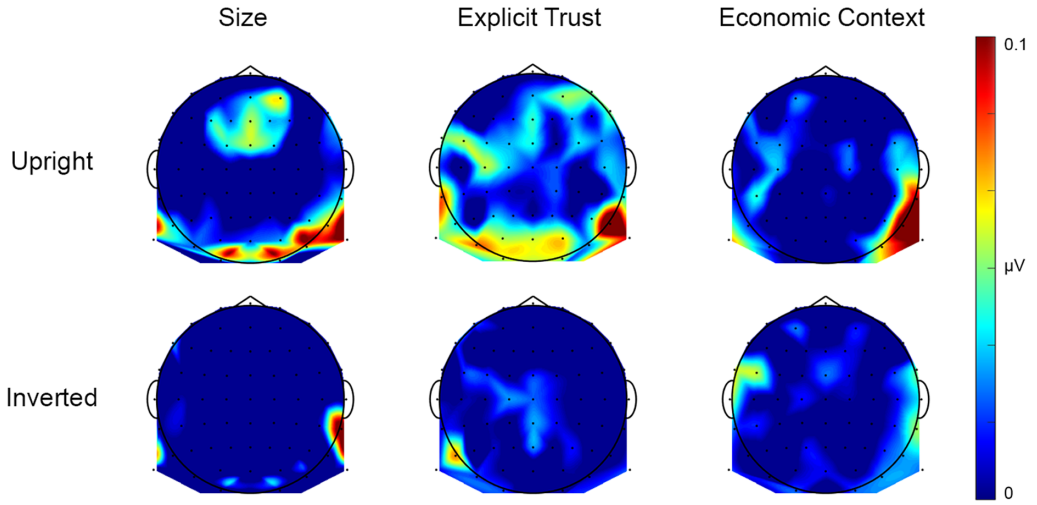


Figure 4. Scalp topographies at the 1 Hz oddball frequency for each task instruction condition in the second block. The conditions differed in instructions given to participants in the second block. In the size condition, participants were asked to judge the size of the faces and determine whether the faces were small or overall. In the explicit trust condition, participants were asked to judge the trustworthiness of the face and determine whether the faces were trustworthy or untrustworthy overall. In the economic context condition, participants were asked to determine whether they would lend money to the faces shown on the screen. Of each set, the top row represents the upright condition and the bottom row represents the inverted condition. The color bar represents the magnitude of neural activation. Hotter colors represent stronger activation.

conditions. Post hoc comparisons of size versus explicit trust and economic context versus explicit trust also revealed posterior odds of 0.19 and 0.26 respectively, indicating moderate evidence in favor of the null hypothesis. The comparison between the size and economic context conditions revealed posterior odds of 0.87, indicating anecdotal evidence in favor of the null hypothesis. See Supplementary Tables S3 and S4 for the full Bayes factor table. Taken together, the Bayesian analyses indicated moderate to decisive evidence that there was no difference across task conditions.

Complementary to the Bayesian analyses, we also used classical statistics to examine the effect of task instruction on the strength of the neural trustworthiness discrimination response. A three-way mixed ANOVA was run, with block (1 or 2) and face orientation (upright or inverted) as the within-subject factors and condition (size, explicit trust, or economic context) as the between-subjects factor. A significant main effect of face orientation, $F(1, 83) = 53.22$, $p < 0.001$ (see Figure 4 for scalp topographies for both orientations for each of the three task conditions and Figure 5 for violin plots of individual data) was found, reflecting stronger responses for upright compared to inverted faces. The main effects of block, $F(1, 83) = 0.33$, $p = 0.57$, $\eta^2 = 0.00$, and condition, $F(1, 83) = 1.13$, $p = 0.33$, $\eta^2 = .01$, were nonsignificant (see Supplementary Figure S3 for violin plots of individual data for each block and condition). Moreover, the two-way interaction effects of block and condition $F(1, 83) = 1.44$, $p = 0.24$, $\eta^2 = 0.00$, orientation and condition, $F(1, 83) = 0.00$, $p = 1.00$, $\eta^2 = 0.00$, and block and orientation $F(1, 83) = 0.48$, $p = 0.49$, $\eta^2 = 0.00$, as well as the three-way interaction effect $F(1, 83) = 0.55$, $p = 0.58$, $\eta^2 = 0.00$, were all nonsignificant. Follow-up contrasts showed no difference between the within-subject factors for any group: size (block 1) versus size (block 2): $t(28) = 0.51$, $p = 0.62$, $d = 0.09$, size (block 1) versus explicit trust (block 2) $t(28) = 0.85$, $p = 0.40$, $d = 0.16$, and size (block 1) versus economic context (block 2), $t(27) = -1.39$, $p = 0.18$, $d = -0.26$. We did not find any evidence that the mean strength of the neural trustworthiness discrimination response differed by task condition.

## Test-retest reliability

Finally, we were also interested in understanding whether it is possible to use the FPVS response as an index of individual participant sensitivity to face trustworthiness. Therefore we examined whether the FPVS responses were reliable across participants. Reliability is not often measured using FPVS (or indeed, any EEG paradigm), but it is an important and useful metric when developing a measure of individual differences (Stacchi, Liu-Shuang, Ramon, & Caldara, 2019). A reliable measure provides greater confidence in the results being stable, which allows us to determine whether the measure can be useful at the individual level. We calculated test-retest reliability at the individual participant level by correlating the oddball response in the upright condition between the first and second FPVS blocks, which were identical in terms of

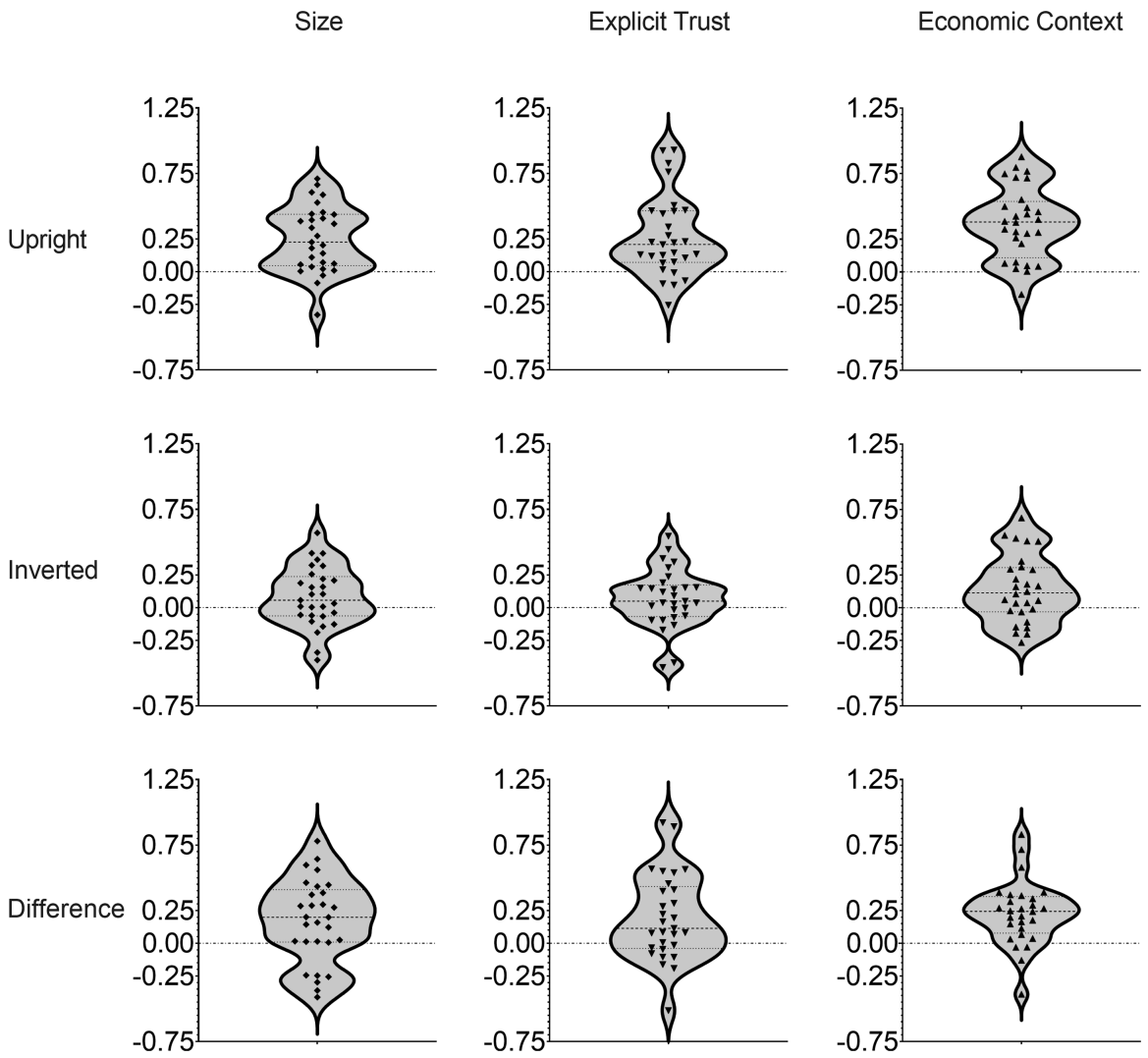


Figure 5. Violin plots with individual summed oddball response data points for each judgement condition in the second block, for the upright and inverted conditions, and their difference (inverted subtracted from upright). All values are in microvolts (μV).

stimulus presentation. To control for individual general responsiveness to EEG or FPVS or faces in general, we additionally calculated the reliability of the upright condition after residualizing for the inverted condition (DeGutis, Wilmer, Mercado, & Cohan, 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 first calculated the reliability for participants using residuals in the size condition only, because this condition represents the clearest measure of reliability given that instructions were identical across blocks. Because the data were suggested to be normally distributed, the Pearson correlation was used to estimate reliability. There was moderate test-retest reliability: $r = 0.47$, [0.13, 0.71], $p = 0.01$, $N = 29$.

Because there was no effect of task instructions on the neural response, we also then correlated the residuals for all participants across block 1 and 2 to take advantage of the larger sample size. This measure also showed moderate test-retest reliability overall: $r = 0.50$, [0.31, 0.64], $p < 0.001$, $N = 86$. Note that similarity between the signal at time one and time two is likely underestimated here as we also capture measurement error in our measure of reliability. Corresponding scatterplots are shown in Figure 6.

## Manipulation checks

### *Stimuli*

As a manipulation check, we examined whether the trustworthy and untrustworthy face stimuli differed in trustworthiness ratings. As expected,

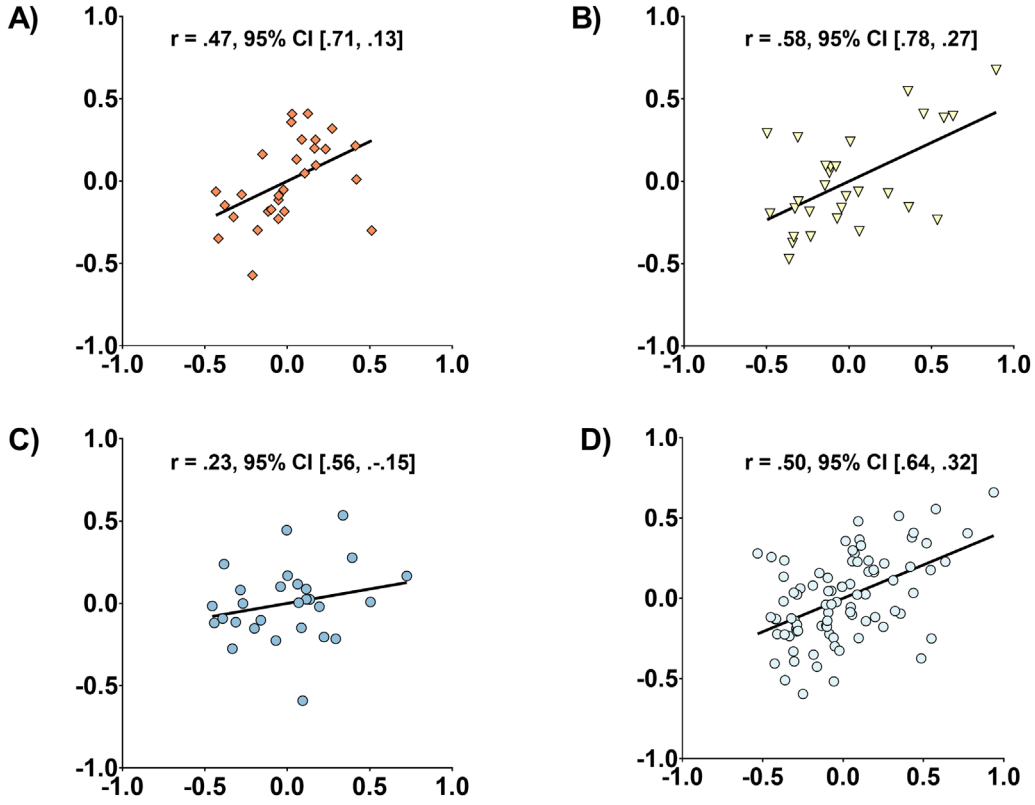


Figure 6. Scatter plots for the residuals (upright controlling for inverted) of the (A) size condition ($N = 29$), (B) trust condition ($N = 29$), (C) economic context condition ($N = 28$), and (D) all participants ($N = 86$). Block 1 (always size) is on the x axis, and block 2 (size, trust, or economic context) is on the y axis. All units are in microvolts (μV).

trustworthy faces ($M = 6.16$, $SD = 0.56$) were rated as more trustworthy than the untrustworthy faces ($M = 3.56$, $SD = 0.93$) in the behavioral ratings task that followed the FPVS experiment, confirming that the manipulation of the stimuli was successful (Supplementary Figure S4).

### *Task*

During the FPVS task, participants made judgements on face size, face trustworthiness, or propensity to lend money, based on the judgement condition in the second block. Accuracy was measured by the percentage of correctly reported sequences. Because the size of the faces in any given sequence were either mainly larger or smaller compared to other sequences (the upper and lower bounds of the jitter was changed for each sequence), participants needed to correctly name the size of the sequence for a correct answer in the size condition. Because the faces were also either mainly trustworthy or untrustworthy, participants needed to correctly identify the trustworthiness of the faces in the trustworthy or economic lending conditions (e.g., in an untrustworthy oddball sequence, given that most faces should look trustworthy, "trustworthy" or "yes," were the correct answers respectively). Importantly, accuracy was high, and similar for the size ($M = 71.21$, $SD = 22.55$), trustworthiness ($M = 68.97$, $SD = 13.65$), and economic ($M = 72.32$, $SD = 15.66$) conditions, $F(2, 83) = 0.27$, $p = 0.77$. A main effect of orientation was found, indicating that accuracy was higher for upright ($M = 72.67$, $SD = 17.25$) compared to inverted ($M = 68.37$, $SD = 21.41$) face sequences, $F(2, 83) = 4.65$, $p = 0.03$. There was no interaction $F(2, 83) = 0.61$, $p = 0.55$.

## Discussion

Our aim was to determine to what extent trustworthiness processing can be shaped by the task context, to evaluate whether trust perception is mandatory. Specifically, we aimed to test whether the strength of the neural trustworthiness discrimination response was modulated by task instruction. The design of our study was such that in the first block, all participants judged the size of faces varying on trustworthiness, as a measure of an implicit brain response to facial trust. In the second block, participants viewed the same faces, across three conditions which varied in task instruction (either

judging the faces on size again, explicit trustworthiness judgements, or choosing whether they would lend money to each person, to make trustworthiness a salient context). Critically, because the face sequences shown in each condition were identical, any differences in neural response between conditions could only be attributable to differences in task instructions given in the second block (and not differences between stimuli). Bayesian analyses found moderate evidence that there was no effect of task instruction on the neural trustworthiness discrimination response, suggesting that, at least when measured at this neural level, trustworthiness discrimination is mandatory and does not require directed instruction. The results also point to the remarkable robustness of trustworthiness processing from faces, which appears to occur from faces viewed extremely rapidly, even when participants are not tasked to judge trust.

Our results are in line with previous studies finding that trustworthiness perception can occur rapidly (Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006), unconsciously (Freeman et al., 2014; Stewart et al., 2012), and implicitly (Klapper et al., 2016; Swe et al., 2020), other important facets of automaticity (Palermo & Rhodes, 2007). Taken together, these results provide support for the idea that the ability to discriminate between trustworthy and untrustworthy individuals can manifest as an automatic process, perhaps as a by-product of face perception processes used in threat detection (Oosterhof & Todorov, 2008, Zebrowitz, 2004) or because of their social importance (Siddique et al., 2022; Sutherland et al., 2020).

It is important to understand which brain regions are involved in trustworthiness discrimination measured with FPVS. Our FPVS paradigm is unlikely to have directly measured responses from the amygdala or the superior temporal sulcus, areas which are typically regions of interest considered in fMRI studies on trustworthiness processing (Santos, Almeida, Oliveiros, & Castelo, 2016). Instead, previous studies have suggested that FPVS strongly represents activity from the fusiform face area (Jonas, Brissart, Hossu, Colnat-Coulbois, Vignal, Rossion, & Maillard, 2018; Rossion, Retter, & Liu-Shuang, 2020), which is a brain region predominantly responsible for extracting perceptual representations of faces for the purposes of detection and recognition (Kanwisher & Yovel, 2006). The neural trustworthiness discrimination response found here may thus primarily reflect early face-sensitive perceptual processes that are unaffected by task instruction. Our findings are also consistent with the recent claim that trustworthiness processing can be subserved by face-selective brain regions normally associated with "core" face processing (Mattavelli et al., 2012).

Our main result, that task instruction does not modulate trustworthiness discrimination responses, differs from the results from an FPVS study investigating face identity discrimination (Yan et al., 2019). In this study, the authors found that neural responses to changes in face identity were stronger when participants were given a face-related task (responding to face gender changes) compared to a non-face-related control task (responding to fixation cross changes), suggesting that task instruction can modulate the neural response to face identity. There are several reasons why our results here may differ from FPVS research using face identity. Of course, this difference may be due to methodological differences, as Yan et al. (2019) compared face and non-face tasks (a fixation cross change), whereas here the tasks always involved face-related judgement. Additionally, the task in Yan et al. (2019) involved judgement of a specific stimulus, whereas the tasks in our study involved paying attention to all the faces in the sequence. However, participants found the current task straightforward, and accuracy was good.

Alternatively, we may not expect the same results if face identity processing and trustworthiness processing reflect distinct face perception processes. Indeed, individuals with prosopagnosia, who show impaired face identity processing, can still make typical trustworthiness impressions from faces, suggesting that the mechanisms involved in face identity processing and face trustworthiness processing can be dissociated (Todorov & Duchaine, 2008). Moreover, a recent twin study failed to find a relationship between individual differences in trustworthiness impressions and facial identity recognition ability (Sutherland et al., 2020). Future studies may wish to directly compare trustworthiness and identity processing, as there is currently great interest in understanding how trustworthiness processing in faces overlaps (or not) with other components of face perception (Sutherland et al., 2020; Todorov & Duchaine, 2008). This question is key given that FPVS responses to face identity and face trust both seem to reflect processing in "core" face regions.

Importantly, it is unlikely that our FPVS results can be explained by low-level retinotopic effects, as we jittered the size of the face images on each presentation. Furthermore, when the faces were inverted (a manipulation known to disrupt normal face processing; Rhodes, Brake, & Atkinson, 1993), we did not observe significant neural responses at the fundamental frequency for inverted faces, and weaker responses at the harmonics. Instead, these findings suggest that there is potentially a strong stimulus-driven component in processing the trustworthiness of faces, such that the processing can be mandatory and independent of top-down factors such as internal volition, goals, and motivation. This idea is also supported by an fMRI study using an implicit task which found that properties of the face stimuli predicted amygdala responses to

changes in facial trustworthiness more so than did variability in individual participants' judgements, indicating a strong stimulus-driven component to trustworthiness processing (Engell, Haxby, & Todorov, 2007). An interesting question for future research is to understand precisely which components (mechanistic or neural) are more or less stimulus driven.

Although the inverted face condition helps mitigate against low-level confounds, it is important to note that the FPVS oddball paradigm inherently indexes the visual discrimination response to the physical features of different stimuli categories. A number of face attributes have been shown to correlate with trustworthiness impressions, such as larger eyes, higher eyebrows and so on (Oosterhof & Todorov, 2008). Indeed, the stimuli used here likely capture a mixture of these key physical differences integral to visual trustworthiness processing, which then drive the FPVS oddball response. However, it is unlikely that any one physical attribute is driving the oddball response, as the original stimuli were created to vary maximally on the trustworthiness dimension (Todorov et al., 2013). Thus the oddball response likely reflects discrimination particular to the trustworthiness dimension, based on the holistic combination of multiple cues involved (Vernon, Sutherland, Young, & Hartley, 2014).

Finally, our finding of a significant neural trustworthiness discrimination response at the group level for upright but not inverted faces, provides converging evidence with Swe et al. (2020) and Verosky et al. (2020) that trustworthiness processing can be indexed using FPVS. Building on Swe et al. (2020), we also found that this marker is reliable and shows variability across participants, suggesting that this technique could be used in future to measure a face-selective neural marker for trustworthiness perception. These findings provide further evidence that FPVS can be a useful tool in investigating individual differences in trustworthiness processing, a topic which has recently been gaining interest (Hehman, Sutherland, Flake, & Slepian, 2017; Sutherland et al., 2020) but is still not well understood.

## Limitations and future research directions

One limitation to this study is that we used tightly controlled, computer-generated stimuli. The use of controlled stimuli ensures that trustworthiness is the strongest dimension in which the faces vary, while reducing the influence of higher-level confounding factors such as emotion and identity, as well as low-level confounding factors such as color and contrast. However, the faces people see in everyday life are much more varied, and many other cues are used to infer trustworthiness (Sutherland et al., 2013; Vernon et al., 2014). Although the current study is focused on stimuli control, an important avenue for future studies could be to use more diverse stimuli. There are two reasons for future studies to use different sets of stimuli. First, it is important for effects to be systematically retested using a range of different stimulus sets to make sure that the effect generalizes across different types of face stimuli (and not dependent on a specific set of stimuli). Second, it is possible that neural responses to other trustworthiness cues, such as gender stereotypes (Sutherland et al., 2015) or cultural differences (Jones et al., 2021; Sutherland, Liu, Zhang, Chu, Oldmeadow, & Young, 2018) would be more likely to be modulated by task instruction, as previous studies have shown modulatory effects of stereotypes on neural activity in person perception areas (Quadflieg, Flannigan, Waiter, Rossion, Wig, Turk, & Macrae, 2011). Additionally, it may be the case that some cues may be processed more implicitly than others. For example, there is evidence to support implicit processing of facial emotion (Dzhelyova, Jacques, & Rossion, 2017), whereas it is less clear whether face gender can be processed implicitly or not (Wiese, Schweinberger, & Neumann, 2008). Future studies could also investigate task instruction modulation of trustworthiness perception using more naturalistic faces with more variation between them, compare the effects of task instruction when using controlled stimuli versus naturalistic and varied stimuli or investigate the different facial attributes that contribute to the trustworthiness dimension in isolation.

Our current results raise the interesting further question of the malleability of trustworthiness processing at different time points. This FPVS paradigm was not designed to capture time course information (see Rossion, Retter, & Liu-Shuang, 2020 for a clear discussion of this point). Indeed, event-related potentials studies looking at explicit trustworthiness judgements have found that untrustworthy faces are processed at many time points, with early processing argued to reflect integrated task and stimulus driven processes and later processing (250 ms/waveforms) possibly reflecting involvement of cognitive and motivational factors (Marzi et al., 2014; Yang, Qi, Ding, & Song, 2011). Other evidence shows that trustworthiness impressions can be affected both by perceptual features and memory retrieval, and that each strategy involves separate neural processes (Rudoy & Paller, 2009). Similarly, behavioral studies are starting to show the importance of the context for shifting threat processing, including altering trustworthiness impressions from the face (Brambilla et al., 2018). Given this work, it is possible that more elaborate social cognitive processing of trustworthiness, presumably occurring at later stages of processing, may be less

automatic or more task dependent. This possibility remains to be investigated.

In this study we investigated task instruction and created conditions involving different judgement instructions. Future studies could instead investigate whether attentional capacity can modulate trustworthiness perception. Although our current study likely shifted the focus of attention on the faces by instructing individuals to judge them in different ways (e.g., attention may shift to the silhouette of the face in the size condition, and to the internal features in the explicit trustworthiness condition), it is also possible that the ability to find differences in the neural trustworthiness discrimination response is restricted by the strong response to faces (i.e., the measured response may already be at its strongest). Indeed, the trust response measured here was highly stable and robust, which is striking considering that the face images are presented at a very rapid frequency (1 Hz/6 Hz). Thus future studies may find it useful to parametrically vary attentional resources instead. For example, attentional load could be manipulated within an FPVS paradigm through distractor tasks, such as the *n*-back task incorporated into a fixation cross task where participants need to respond when the fixation cross changes to a shape that occurred two shape changes previously. Additionally, a dual task approach, which has previously been used in natural scene categorization (Li, VanRullen, Koch, & Perona, 2002) could be used, whereby an attentionally demanding central task is performed concurrently with a peripheral FPVS task. If the neural trustworthiness discrimination response is weaker during the distractor task or dual task condition, then it would suggest that trustworthiness perception can be modulated by attention; conversely, if trustworthiness responses are unaffected, results may suggest that trustworthiness perception is capacity free.

Finally, research could investigate whether task instruction can modulate other face perception domains using FPVS. As noted, previous research has shown modulation effects of task instruction for face identity (Yan et al., 2019), whereas here we do not find any effects of task instruction for face trustworthiness. It would be useful to investigate other domains to determine whether there are any systematic differences between face perception processes that make them susceptible or resistant to modulation by task instruction at the neural level. In the case of face emotion, for example, some studies have shown evidence of neural responses to facial emotion being modulated by task instruction using fMRI (Lange, Williams, Young, Bullmore, Brammer, Williams, Gray, Phillips, 2003; Narumoto, Okada, Sadato, Fukui, & Yonekura, 2001). Thus it is plausible that similar results may be found for face emotion processing using FPVS as well.

## Conclusions

In conclusion, task instruction was not found to modulate trustworthiness perception at the neural level as measured by FPVS, suggesting trustworthiness processing has a strong stimulus-driven component and pointing to the mandatory nature of trustworthiness discrimination. Our results also provide further evidence toward the robust nature of FPVS and lend support toward the utility of this technique to investigate higher-level (face) perception at both the group and individual level.

*Keywords: trustworthiness impressions, EEG, fast periodic visual stimulation, FPVS, implicit perception, SSVEP, task-set, task instruction, automatic, mandatory*

## Acknowledgments

The authors 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.

Supported by an RTP scholarship from the University of Western Australia to DS, an Australian Research Council (ARC) Discovery Early Career Research Award to CS (DE190101043), and ARC Discovery Projects to CS and RP (DP170104602) and RP (DP140101743).

Commercial relationships: none.
Corresponding author: Derek Swe.
Email: derek.swe@research.uwa.edu.au.
Address: School of Psychological Science, The University of Western Australia, WA 6009, Australia.

## References

Beck, A. A., Rossion, B., & Samson, D. (2017). An objective neural signature of rapid perspective taking. *Social Cognitive and Affective Neuroscience, 13*(1), 72–79, https://doi.org/10.1093/scan/nsx135.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current directions in psychological science, 20*(5), 286–290, https://doi.org/10.1177/0963721411422522.

Brambilla, M., Biella, M., & Freeman, J. B. (2018). The influence of visual context on the evaluation of facial trustworthiness. *Journal of Experimental Social Psychology, 78*, 34–42, https://doi.org/10.1016/j.jesp.2018.04.011.

Brambilla, M., Masi, M., Mattavelli, S., & Biella, M. (2021). Faces and sounds becoming one: Cross-modal integration of facial and auditory cues in judging trustworthiness. *Social Cognition, 39*(3), 315–327, https://doi.org/10.1521/soco.2021.39.3.315.

Collova, J. R., Sutherland, C. A., & 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, https://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., Jacques, C., & Rossion, B. (2017). 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), 4106–4123, https://doi.org/10.1093/cercor/bhw223.

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., & 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.

Eggleston, A., Flavell, J. C., Tipper, S. P., Cook, R., & Over, H. (2021). Culturally learned first impressions occur rapidly and automatically and emerge early in development. *Developmental Science, 24*(2), e13021, https://doi.org/10.1111/desc.13021.

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), 1508–1519, https://doi.org/10.1162/jocn.2007.19.9.1508.

Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). Looking the part: Social status cues shape race perception. *PloS one, 6*(9), e25107, https://doi.org/10.1371/journal.pone.0025107.

Freeman, J. B., Stolier, R. M., Ingbretsen, Z. A., & Hehman, E. A. (2014). Amygdala responsivity to high-level social information from unseen faces. *Journal of Neuroscience, 34*(32), 10573–10581, https://doi.org/10.1523/JNEUROSCI.5063-13.2014.

Goss-Sampson, M., van Doorn, J., & Wagenmakers, E. (2020). *Bayesian inference in JASP: A guide for students*. University of Amsterdam: JASP team 46.

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*, 405–416, 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, http://doi.org/10.1348/000712607X199981.

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.

Jaeger, B., Sleegers, W. W., Evans, A. M., Stel, M., & van Beest, I. (2019). The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. *Journal of Economic Psychology, 75*, 102125, https://doi.org/10.1016/j.joep.2018.11.004.

Jaeger, B., Todorov, A. T., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology, 90*, 104004, https://doi.org/10.1016/j.jesp.2020.104004.

Jeffreys, H. (1961). *The theory of probability*. Oxford: Oxford University Press.

Jonas, J., Brissart, H., Hossu, G., Colnat-Coulbois, S., Vignal, J.-P., Rossion, B., . . . Maillard, L. (2018). A face identity hallucination (palinopsia) generated by intracerebral stimulation of the face-selective right lateral fusiform cortex. *Cortex, 99*, 296–310, https://doi.org/10.1016/j.cortex.2017.11.022.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., . . . Foroni, F. (2021). To which world regions does the valencedominance model of social perception apply? *Nature Human Behaviour, 5*(1), 159–169, https://doi.org/10.1038/s41562-020-01007-2.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1476), 2109–2128, https://doi.org/10.1098/rstb.2006.1934.

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, http://dx.doi.org/10.1037/pspa0000090.

Lange, K., Williams, L. M., Young, A. W., Bullmore, E. T., Brammer, M. J., Williams, S. C., . . . Phillips, M. L. (2003). Task instructions

modulate neural responses to fearful facial expressions. *Biological Psychiatry, 53*(3), 226–232, https://doi.org/10.1016/S0006-3223(02)01455-5.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences, 99*(14), 9596–9601, https://doi.org/10.1073/pnas.092277599.

Linke, L., Saribay, S. A., & Kleisner, K. (2016). Perceived trustworthiness is associated with position in a corporate hierarchy. *Personality and Individual Differences, 99*, 22–27, https://doi.org/10.1016/j.paid.2016.04.076.

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.

Marzi, T., Righi, S., Ottonello, S., Cincotta, M., & Viggiano, M. P. (2014). Trust at first sight: evidence from ERPs. *Social Cognitive and Affective Neuroscience, 9*(1), 63–72, https://doi.org/10.1093/scan/nss102.

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), 2205–2211, https://doi.org/10.1016/j.neuropsychologia.2012.05.024.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3, https://doi.org/10.1037/1076-8971.7.1.3.

Moors, A., & De Houwer, J. (2012). How to define and examine implicit processes. *Psychology of Science: Implicit and Explicit Processes*, 183–198,.

Narumoto, J., Okada, T., Sadato, N., Fukui, K., & Yonekura, Y. (2001). Attention to emotion modulates fMRI activity in human right superior temporal sulcus. *Cognitive Brain Research, 12*(2), 225–231, https://doi.org/10.1016/S0926-6410(01)00053-2.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences, 18*(11), 566–570, https://doi.org/10.1016/j.tics.2014.09.007.

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), 315–324, https://doi.org/10.1016/j.jesp.2009.12.002.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087–11092, https://doi.org/10.1073/pnas.0805664105.

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia, 45*(1), 75–92, https://doi.org/10.1016/j.neuropsychologia.2006.04.025.

Quadflieg, S., Flannigan, N., Waiter, G. D., Rossion, B., Wig, G. S., Turk, D. J., … Macrae, C. N. (2011). Stereotype-based modulation of person perception. *NeuroImage, 57*(2), 549–557, https://doi.org/10.1016/j.neuroimage.2011.05.004.

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.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS one, 7*(3), e34293, https://doi.org/10.1371/journal.pone.0034293.

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), 1599–1621, 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), 1585–1600, https://doi.org/10.1016/j.neuroimage.2012.08.033.

Rossion, B., Retter, T. L., & Liu-Shuang, J. (2020). Understanding human individuation of unfamiliar faces with oddball fast periodic visual stimulation and electroencephalography. *Eur J Neurosci, 52*(10), 4283–4344, https://doi.org/10.1111/ejn.14865.

Rudoy, J. D., & Paller, K. A. (2009). Who can you trust? Behavioral and neural differences between perceptual and memory-based influences. *Frontiers in Human Neuroscience, 3*, 16, https://doi.org/10.3389/neuro.09.016.2009.

Santos, I., & Young, A. (2005). Exploring the perception of social characteristics in faces using

the isolation effect. *Visual Cognition, 12*(1), 213–247, https://doi.org/10.1080/13506280444000102.

Santos, S., Almeida, I., Oliveiros, B., & Castelo-Branco, M. (2016). The role of the amygdala in facial trustworthiness processing: A systematic review and meta-analyses of fMRI studies. *PloS one, 11*(11), e0167276.

Siddique, S., Jeffery, L., Palermo, R., Collova, J., & Sutherland, C.A.M. (In Press). Children's dynamic use of face- and behaviour-based cues in an economic trust game. *Developmental Psychology.*

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

Stacchi, L., Liu-Shuang, J., Ramon, M., & Caldara, R. (2019). Reliability of individual differences in neural face identity discrimination. *NeuroImage, 189*, 468–475, https://doi.org/10.1016/j.neuroimage.2019.01.023.

Stewart, L. H., Ajina, S., Getov, S., Bahrami, B., Todorov, A., & Rees, G. (2012). Unconscious evaluation of faces on social dimensions. *Journal of Experimental Psychology: General, 141*(4), 715, http://doi.org/10.1037/a0027950.

Sutherland, C. A., Burton, N. S., Wilmer, J. B., Blokland, G. A., Germine, L., Palermo, R., … Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences, 117*(19), 10218–10224, https://doi.org/10.1073/pnas.1920131117.

Sutherland, C. A., Liu, X., Zhang, L., Chu, Y., Oldmeadow, J. A., & Young, A. W. (2018). Facial first impressions across culture: Data-driven modeling of Chinese and British perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin, 44*(4), 521–537, https://doi.org/10.1177/0146167217744194.

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), 105–118, 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.

Swe, D. C., Palermo, R., Gwinn, O. S., Rhodes, G., Neumann, M., Payart, S., … Sutherland, C. A. (2020). An objective and reliable electrophysiological marker for implicit trustworthiness perception. *Social Cognitive and Affective Neuroscience, 15*(3), 337–346, https://doi.org/10.1093/scan/nsaa043.

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, https://psycnet.apa.org/doi/10.1037/a0032335.

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive neuropsychology, 25*(3), 395–410, https://doi.org/10.1080/02643290802044996.

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*(1), 519–545, https://doi.org/10.1146/annurev-psych-113011-143831.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833, https://doi.org/10.1521/soco.2009.27.6.813.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*(12), 455–460, https://doi.org/10.1016/j.tics.2008.10.001.

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), 2766–2781.

Tracy, R. E., Wilson, J. P., Slepian, M. L., & Young, S. G. (2020). Facial trustworthiness predicts ingroup inclusion decisions. *Journal of Experimental Social Psychology, 91*, 104047, https://doi.org/10.1016/j.jesp.2020.104047.

Van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796–803, 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.

Verosky, S. C., Zoner, K. A., Marble, C. W., Sammon, M. M., & Babarinsa, C. O. (2020). Implicit responses to face trustworthiness measured with fast periodic visual stimulation. *Journal of Vision, 20*(7), 29–29, https://doi.org/10.1167/jov.20.7.29.

Wiese, H., Schweinberger, S. R., & Neumann, M. F. (2008). Perceiving age and gender in unfamiliar faces: Brain potential evidence for implicit and explicit person categorization. *Psychophysiology, 45*(6), 957–969, https://doi.org/10.1111/j.1469-8986.2008.00707.x.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598, 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, http://dx.doi.org/10.1038/nn816.

Yan, X., Liu-Shuang, J., & Rossion, B. (2019). Effect of face-related task on rapid individual face discrimination. *Neuropsychologia, 129*, 236–245, https://doi.org/10.1016/j.neuropsychologia.2019.04.002.

Yang, D., Qi, S., Ding, C., & Song, Y. (2011). An ERP study on the time course of facial trustworthiness appraisal. *Neuroscience Letters, 496*(3), 147–151, 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), 2256–2262, https://doi.org/10.1016/j.cub.2005.10.072.

Zebrowitz, L. A. (2004). The origin of first impressions. *Journal of Cultural and Evolutionary Psychology, 2*(1-2), 93–108, http://dx.doi.org/10.1556/JCEP.2.2004.1-2.6.

Zebrowitz, L. A. (2017). First impressions from faces. *Current Directions in Psychological Science, 26*(3), 237–242, https://doi.org/10.1177/0963721416683996.