Visual Cognition

ISSN: 1350-6285 (Print) 1464-0716 (Online) Journal homepage: https://www.tandfonline.com/loi/pvis20

High test-retest reliability of a neural index of rapid automatic discrimination of unfamiliar individual faces

Milena Dzhelyova, Corentin Jacques, Giulia Dormal, Caroline Michel, Christine Schiltz & Bruno Rossion

To cite this article: Milena Dzhelyova, Corentin Jacques, Giulia Dormal, Caroline Michel, Christine Schiltz & Bruno Rossion (2019): High test-retest reliability of a neural index of rapid automatic discrimination of unfamiliar individual faces, Visual Cognition, DOI: 10.1080/13506285.2019.1616639

To link to this article: https://doi.org/10.1080/13506285.2019.1616639

Published online: 19 Jun 2019.

VISUAL COGNITION
https://doi.org/10.1080/13506285.2019.1616639

Routledge
Taylor & Francis Group

High test-retest reliability of a neural index of rapid automatic discrimination of unfamiliar individual faces

Milena Dzhelyova[a,b], Corentin Jacques[a,c], Giulia Dormal[a], Caroline Michel[a], Christine Schiltz[b] and Bruno Rossion[a,d,e]

[a]Psychological Sciences Research Institute and Institute of Neuroscience, Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium; [b]Cognitive Science and Assessment Institute (COSA), University of Luxembourg, Luxembourg; [c]Center for Developmental Psychiatry, Department of Neurosciences, KU Leuven, Leuven, Belgium; [d]Université de Lorraine, CNRS, CRAN, Nancy, France; [e]Université de Lorraine, CHRU-Nancy, Service de Neurologie, Nancy, France

ABSTRACT

An important aspect of human individual face recognition is the ability to discriminate unfamiliar individual. Since many general processes contribute to explicit behavioural performance in individual face discrimination tasks, isolating a measure of unfamiliar individual face discrimination ability in humans is challenging. In recent years, a fast periodic visual stimulation approach (FPVS) has provided objective (frequency-locked) implicit electrophysiological indices of individual face discrimination that are highly sensitive at the individual level within a few minutes of testing. Here we evaluate the test-retest reliability of this response across scalp electroencephalographic (EEG) recording sessions separated by more than two months, in the same 30 individuals. We found no test-retest difference overall across sessions in terms of amplitude and spatial distribution of the EEG individual face discrimination response. Moreover, with only 4 stimulation sequences corresponding to 4 min of recordings per session, the individual face discrimination response was highly reliable in terms of amplitude, spatial distribution, and shape. Together with previous observations, these results strengthen the diagnostic value of FPVS-EEG as an objective and rapid flag for specific difficulties at individual face recognition in the human population.

ARTICLE HISTORY
Received 8 August 2018
Accepted 19 March 2019

KEYWORDS
Fast periodic visual stimulation; EEG; individual face discrimination; reliability

Introduction

In the human species, recognizing individual people by their faces is critical for social interactions. A key aspect of this individual face recognition (IFR) function is the ability to discriminate different but highly similar visual patterns, i.e., individual faces, even when these faces have not been encoded before in memory, i.e., are unfamiliar to us.

Measuring unfamiliar face discrimination ability in humans is challenging. Early behavioural studies required human participants to match images of the same face identity against distractor individual faces (e.g., Benton & Van Allen, 1968; De Renzi & Spinnler, 1966). Since then, such tasks have been used in countless studies in order to understand the nature of IFR. Depending on the type of stimuli used, the degree of generalization required by the matching task (i.e., matching the exact same image or images varying in size, head orientation or lighting direction), the number and variability of distractors, as well as the task instructions, the overall performance of healthy human adults at unfamiliar face discrimination tasks varies from near ceiling to 80%-70% accuracy, with substantial variability in performance across individuals in some studies (e.g., Bowles et al., 2009; Busigny & Rossion, 2010; Estudillo & Bindemann, 2014; Herzmann, Danthiir, Schacht, Sommer, & Wilhelm, 2008; Sergent, 1984; Bruce et al., 1999; Bruce, Henderson, Newman, & Burton, 2001; Megreya & Burton, 2006; Rossion & Michel, 2018). Although this performance has sometimes been described as being seriously limited or even "poor" (e.g., Megreya & Burton, 2006; Young & Burton, 2018), it is in fact remarkable compared to performance at similar tasks of young children, brain-damaged patients with prosopagnosia or other animal species such as macaque monkeys, suggesting that neurotypical humans have a genuine expertise at

CONTACT Bruno Rossion bruno.rossion@univ-lorraine.fr CRAN, UMR 7039, CNRS - Université de Lorraine, Pavillon Krug (1er étage - entrée CC-1), Hôpital Central, CHRU Nancy - University Hospital of Nancy, 29 Avenue du Maréchal de Lattre de Tassigny, 54000 Nancy, France

discrimination of individual unfamiliar faces (Rossion, 2018; Rossion & Taubert, in press).

Nevertheless, human performance at explicit behavioural tasks can be influenced by many factors beyond the individual face discrimination function, such as task understanding, motivation, as well as attentional and decisional processes. This makes the comparison of individual face discrimination ability between individuals, as well as the assessment of ability across development and in clinical populations particularly challenging. Moreover, unfamiliar individual face discrimination performance in explicit behavioural tasks tends to be reflected in two variables, accuracy rates and response times (RTs), which may both carry relevant information for evaluating this function (Rossion & Michel, 2018). Finally, explicit behavioural tasks lack validity at two levels at least. First, since they rely on explicit instructions and behavioural responses, these tasks do not directly measure *automatic* individual face discrimination. This is unfortunate because in real life circumstances, typical human adults discriminate individual faces automatically, i.e., without the intention to do so and without being able to suppress this visual discrimination process. Second, these tasks do not measure *rapid* individual face discrimination, i.e., forcing the visual recognition system to perform this function at a single glance. Indeed, although face stimuli can be presented for a limited time duration in such tasks, this is rarely the case in behavioural studies, in which faces are often presented for very long, or even unlimited, durations (e.g., the Cambridge Face Memory Test, CFMT, Duchaine & Nakayama, 2006). One reason for that is that time pressure in explicit unfamiliar face discrimination tasks can deteriorate behavioural performance even in healthy adult participants (Bindemann, Fysh, Cross, & Watts, 2016; Fysh & Bindemann, 2017) and could even be more problematic when testing children or clinical populations.

To overcome these issues, recent studies have taken advantage of the brain's property to exactly synchronize its electrophysiological activity to the temporal frequency of a stimulus (Adrian & Matthews, 1934; Regan, 1989; see Norcia, Appelbaum, Ales, Cottereau, & Rossion, 2015 for review). Specifically, by coupling fast periodic visual stimulation (FPVS) with human electroencephalography (EEG), one can obtain valid, objective and sensitive measures of unfamiliar individual face discrimination (Alonso-Prieto, Van Belle, Liu-Shuang, Norcia, & Rossion, 2013; Dzhelyova & Rossion, 2014a, 2014b; Liu-Shuang, Norcia, & Rossion, 2014; Rossion, Prieto, Boremanse, Kuefner, & Van Belle, 2012; Rossion & Boremanse, 2011; Xu, Liu-Shuang, Rossion, & Tanaka, 2017). A key paradigm for this endeavour to succeed is based on a periodic *oddball-like* stimulation, in which robust individual face discrimination measures can be obtained in every typical adult tested within a few minutes (Dzhelyova & Rossion, 2014a, 2014b; Liu-Shuang et al., 2014; Liu-Shuang, Torfs, & Rossion, 2016; Xu et al., 2017). In this paradigm, during each stimulation sequence, a randomly selected unfamiliar face identity is presented repeatedly at a periodic rate, usually 6 Hz (i.e., 6 images/second), allowing only a single fixation on each face image. Different face identities are introduced at a lower periodic rate (e.g., 1 change of identity every 5 faces, or 1.2 Hz). While EEG responses recorded at 6 Hz (and harmonics, i.e., 12 Hz, etc.) reflect common visual processing of all visual stimuli, responses at 6 Hz/5 and its specific harmonics (1.2, 2.4 Hz, etc.) can be taken as an index of rapid (i.e., single-glanced) individual face discrimination, even when there is no instruction to explicitly individuate faces (Dzhelyova & Rossion, 2014a, 2014b; Liu-Shuang et al., 2014; Liu-Shuang et al., 2016; Xu et al., 2017). Thanks to the frequency-domain transform of about 60s recording epochs, this individual face discrimination response is objectively identified and quantified (i.e., measured only at an experimentally pre-defined small frequency bin of interest) and sensitive (i.e., associated with a high signal-to-noise ratio, SNR) (see Rossion, 2014).

Importantly, in typical individuals, this individual face discrimination response – which is usually the largest over the right occipito-temporal cortex – resists low-level stimulus manipulation such as substantial changes of stimulus size (Dzhelyova & Rossion, 2014a; Liu-Shuang et al., 2014). However, it is largely decreased following stimulus inversion and contrast-reversal (Liu-Shuang et al., 2014), two manipulations preserving low-level visual cues but severely affecting individual face recognition (Galper, 1970; Yin, 1969). Moreover, this electrophysiological index can be selectively affected in patients with prosopagnosia following brain damage (Liu-Shuang et al., 2016) and in individuals with autism spectrum disorder (ASD, Vettori et al., 2019), showing its relevance for measuring individual face recognition.

Most recently, the FPVS-EEG approach started to be used to characterise interindividual variability in individual face discrimination ability and define its relation to explicit learning tests with unfamiliar faces (Xu et al., 2017). In this context, an outstanding issue is whether the FPVS-EEG measure of individual face discrimination measure is *reliable,* at the group level and most importantly at the individual level. Although reliability is the pedestal of any scientific research, ensuring reproduction of results across laboratories, it is rarely evaluated, especially for neural measures of high-level functions such as individual face discrimination. In Xu et al. (2017), this issue was addressed with a low-density channel system by comparing the stability of the EEG amplitude across individuals tested *during the same recording session*. The present study goes beyond these observations by testing the same individuals in the same brief FPVS-EEG oddball-like paradigm (i.e., about 4 min of data recording) across two different high-density EEG sessions separated by more than two months. In addition, the reliability of the individual face discrimination response is assessed at three levels: (1) in terms of its amplitude at each channel with a fine sampling of electrophysiological activity (i.e., 128 channels), (2) with respect to its spatial topography and (3) in qualitative terms, as characterized by its distribution of amplitude across different harmonics, reflecting the morphology of the neural response. Finally, we compared the between-session to the within-session reliability and estimated the variation in signal-to-noise ratio (SNR) of the discrimination response as a function of the number of sequences presented.

Methods

+Participants

Thirty-two participants (17 females; Mean ± SD age at first recording session = 22.12 ± 2.62) were recruited to take part in the two recording sessions. They were all right-handed, free of neurological or psychiatric problems and had normal or corrected-to-normal vision. All participants provided signed and informed consent and were paid according to their testing time. The study was approved by the Biomedical Ethical Committee of the University of Louvain. Two participants were excluded due to inconsistencies in the procedure of the test-retest session (i.e., different cap sizes were used) and replaced with two participants to reach the initially targeted sample of 30 individuals tested.

Stimuli

Facial stimuli were 25 female and 25 male photographs from the Face Categorization lab database. A detailed description of the images is available in previous studies investigating unfamiliar individual face discrimination (Laguesse, Dormal, Biervoye, Kuefner, & Rossion, 2012; with FPVS: Dzhelyova & Rossion, 2014a, 2014b; Liu-Shuang et al., 2014). All faces were unknown to the participants tested. They were presented at a frontal view with forward eye gaze, masked external features such as ears and hair, and placed against a grey background (Figure 1). Images were resized to 250 pixels height (width = 186 +−11 pixels), corresponding to 8.57 deg x 3.97 deg at an 80 cm distance from the monitor.

Procedure

The experiment consisted of test-retest sessions performed on average 2 months apart (Mean ± *SD* = 66 ± 7.80 days, Figure 1A). Different experimenters tested the participants during the first session but also during the retest session. They were seated comfortably in a dimly lit room 80 cm away from the monitor. They performed only 4 stimulation sequences of about 1 min, which is sufficient to provide robust individual face discrimination measures in single individuals (e.g., Liu-Shuang et al., 2014; Liu-Shuang et al., 2016; Vettori et al., 2019; Xu et al., 2017). In each stimulation sequence, a randomly chosen face identity (either male, for 2 sequences, or female, for 2 sequences) was presented repeatedly at a fast rate of 6 Hz. Stimuli were presented through sinusoidal contrast modulation as in previous studies. They varied randomly in size (80–120% of original size) at each cycle, as also performed in previous studies (e.g., Liu-Shuang et al., 2014; Rossion & Boremanse, 2011) in order to minimize low-level cue repetition effects (see Dzhelyova & Rossion, 2014a for quantification of size change effects on the EEG response). Within a given sequence, different same-sex faces picked randomly among the pool of the remaining 24 faces, appeared as every 5th stimulus (i.e., change of identity frequency 6 Hz/5 = 1.2 Hz,

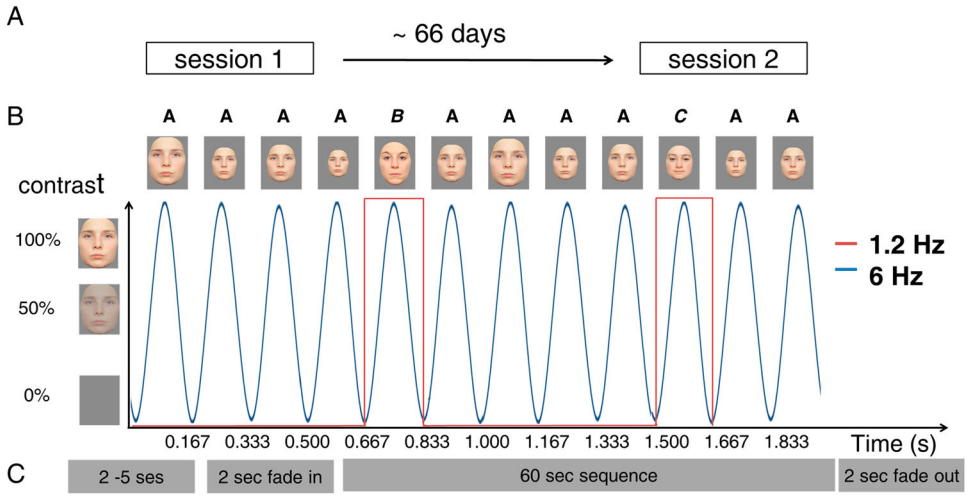

Figure 1. Experimental design. A) Test-retest procedure. B) Example of a FPVS visual stimulation sequence where images of identity A are presented through sinusoidal contrast modulation at 6 Hz and every 5th image is a different identity (B, C etc). Thus, unfamiliar face identity change occurs at 1.2 Hz (6 Hz/5). Image size varies for each cycle. C) Length of an experimental trial.

Figure 1B). Each sequence started with a fixation cross presented for a random period of a 2–5 sec, followed by a 2-sec fade-in interval during which image contrast gradually increased, a 60-sec stimulation sequence and 2-sec fade-out. The fade-in and fade-out were included to avoid abrupt eye movements at stimulation onset and offset. Participants' task was to respond to brief (300 ms) changes in the colour of the fixation cross, and they received no information as to the goal of the study. In session 2, the exact same experiment was performed, using the same pool of individual faces. However, the selection of individual faces was made at random, so that different sequences were actually presented in session 1 and session 2. Participants performed the orthogonal task at ceiling, with no difference in ($ps > 0.40$) accuracy (session 1: $M \pm SEM = 0.96 \pm 0.031$; session 2: 0.98 ± 0.004) or response times ($M \pm SEM$ for the test: 433 ± 17 and retest: 442 ± 7) between the two sessions.

EEG acquisition

EEG was recorded via a BIOSEMI Active two amplifier system (Biosemi, Amsterdam, Netherlands) with 128 Ag/AgCl electrodes inserted in an electrode cap, and sampled at 512 Hz. Electrodes' scalp locations are similar to the standard 10–20 system locations and additional intermediate positions (see Rossion, Torfs, Jacques, & Liu-Shuang, 2015). Electrode positions were not recorded (i.e., digitized) on each participant's head, but measures of the anatomical landmarks used for the essential positioning of the electrodes were taken for each participant (nasion, inion, head circumference) to be preserved as much as possible between the two recording sessions. Eye movements were monitored with four electrodes, one placed at the outer canthi of each eye (HEOG), and one placed above and one below the right eye (VEOG).

EEG preprocessing and frequency domain analysis

All EEG pre-processing steps were carried out with Letswave 6 (http://nocions.github.io/letswave6/), and Matlab (R2012b, Math works) and followed procedures described in detail in previous publications with this approach (see e.g., Liu-Shuang et al., 2016). EEG data were digitally band-pass filtered at 0.10–100 Hz with a Butterworth filter (4^{th} order) and down-sampled to 256 Hz to reduce computation load. Then, it was segmented to include 2 s before and after each sequence (i.e., before the fade-in and after the fade-out of the stimulation), resulting in 68 s segments (−2–66 s). Data from participants who blinked more than 10 times in at least 2 sequences (mean number of blinks across participants = 3.5, $SD = 4.59$), were corrected by means of ICA using the runica algorithm (Bell & Sejnowski, 1995; Makeig, Bell, Jung, & Sejnowski, 1996), as implemented in EEGLAB. This algorithm outputs a square mixing matrix in which the number of components corresponds to the number of channels. For 2 participants, data at both sessions were corrected, while for another 5 participants only

data of one of the sessions was corrected. For each of these participants, only one component representing vertical eye movements was removed. Channels with extreme voltage offset (exceeding ± 100 μV identified by visual examination) were replaced using linear interpolation of the 3 neighbouring channels. Less than 5% of the channels were interpolated per participant, only = 1.1 ± 1.58 (M ± SEM) channels for session 1 and 1.43 ± 1.28 channels for session 2. Given the high SNR of the method due to the concentration of the signal in small frequency bins of interest (Regan, 1989); no further artefact correction was applied. After that, a common average reference computation was applied to all channels for each participant.

Preprocessed data segments were cropped to an integer number of 1.2 Hz cycles, beginning 2 s after the onset of the sequence until approximately 62 s (~60 s, 15149 time samples in total). The first 2 s of each segment (i.e., fade-in) were excluded to avoid any contamination by the initial transient responses. The 4 resulting 60 s segments were averaged in the time-domain to increase the signal to noise ratio (SNR). A Fast Fourier Transform (FFT) was then applied to these averaged segments and normalized amplitude spectra were extracted for all channels (square root of the sum of squares of the real and imaginary parts divided by the number of data points). Thanks to the long time-window, frequency analysis yielded spectra with a high frequency resolution of 0.0166 Hz (1/60), thus increasing SNR (Regan, 1989) and allowing unambiguous identification of the response at the frequency of the change in face identity (1.2 Hz).

The amplitude spectra across participants were averaged for each session and the resulting grand-averaged spectra were averaged together in order to determine the number of significant harmonics independently of the session. The resulting EEG spectrum was averaged across all 128 channels. In order to identify the presence of statistically significant responses at the frequencies of interest and its harmonics, the averaged amplitude spectrum was converted to Z-scores by computing the difference between the amplitude at the frequency of interest and the mean amplitude of 20 surrounding frequency bins divided by the standard deviation of the 20 surrounding bins (see e.g., Liu-Shuang et al., 2016). Only harmonics with significant responses (Z-score > 3.1, $p < 0.001$ one-tailed, i.e., signal > noise) were taken for analysis. Based on this criterion, 6 harmonics (i.e., 1.2, 2.4, 3.6, 4.8, 7.2, 8.4 Hz) were significant and thus included to quantify the individual face discrimination response, while 8 harmonics up to 48 Hz were included for the quantification of the 6 Hz response.

To quantify the individual face discrimination response, the baseline-corrected amplitudes were calculated on individual subjects' spectra in each session by subtracting the mean amplitude of the surrounding bins (until the 10th on each side, excluding the immediately adjacent bin and the bins containing the highest and lowest amplitudes) and summed over the 6 significant harmonics, excluding the 5th harmonic corresponding to the 6 Hz response. The response was quantified over 2 regions of interest (ROIs): in the left occipito-temporal (LOT: electrodes PO7, PO9, PO11, P7, P9) and right occipito-temporal (ROT: PO8, PO10, PO12, P8, P10) sites. These ROIs were defined based on previous studies (Dzhelyova & Rossion, 2014a, 2014b; Liu-Shuang et al., 2014; Liu-Shuang et al., 2016) and visualization of the present data. The 6 Hz response was quantified as the baseline-corrected amplitudes of the summed 8 (up to 48 Hz) significant 6 Hz harmonics over the medial occipital site (MO): POz, POOz, Oz, Oiz, Iz. The summed baseline-corrected amplitudes were averaged across the five electrodes for each ROI.

To evaluate the reliability of the FPVS response across sessions, we performed three types of analyses.

Reliability of EEG amplitude

First, we Pearson-correlated the amplitude across the 30 individual participants between session 1 and session 2 for the LOT/ROT ROIs. We used the summed baseline-corrected amplitudes for the discrimination response either over the left or the right OT region, depending on which hemisphere showed the largest response in a given individual. Hemispheric dominance was determined by the presence of a larger individual face discrimination response at both sessions over a particular hemisphere. When lateralization differed between sessions, the larger difference across the two sessions between the left and right hemisphere was used to determine lateralization. Hemispheric lateralization was confirmed by visualization of scalp topographies.

In addition, we computed the intraclass correlation coefficient (ICC, absolute agreement with one way model) originally applied in interrater reliability

studies, but more recently also applied to assess test-retest reliability of EEG measures (e.g., Cassidy, Robertson, & O'Connell, 2012; Munsters, van Ravenswaaij, van den Boomen, & Kemner, 2019; Clayson & Larson, 2013). Pearson's *r* simply reflects the correlation between testing sessions by ranking the individual participants but does not take into account whether the measurement variance is consistent between the two sessions. One way ICC does not allow systematic variation and thus it provides information about potential mean amplitudes differences of the individual discrimination response across sessions. Thus, high ICC scores are obtained only if there is high consistency in the ranking of the participants and if the mean amplitudes between sessions are similar. *ICC* values can range between 1 and 0, with 1 indicating perfect reliability, i.e., that the signal amplitudes are identical between the sessions within subjects, while a value of 0 indicates no reliability. We used a conservative criterion to determine which ICC values are acceptable: values below 0.40 are considered as poor, values between 0.41 and 0.59 as moderate, values between 0.60 and 0.74 as good, and values at or above 0.75 as excellent.

Replicability of EEG spatial distribution

To evaluate the stability of the scalp topographical response across recording sessions, we ran three different analyses. First, we computed the test-retest correlation between the grand-averaged data of the 30 participants using all 128 channels. In addition to the group level reliability, to estimate the reliability of the individual voltage topography across the two sessions, we computed Pearson correlations, ICC and Cronbach's alpha (e.g., Thigpen, Kappenman, & Keil, 2017) by concatenating the baseline corrected amplitudes from all 128 channels and all 30 participants (i.e., 3840 data points) separately for the two sessions, thus creating a matrix 3840*2 (reliability of individual scalp topographies). We further examined if limiting the reliability analysis to the electrodes within the ROI would increase the Cronbach's alpha as most of the response was observed over occipito-temporal sites (i.e., this will reduce the influence of electrodes not contributing to the response and introducing noise). For Cronbach's alpha, we used the following classification (Hinton, Brownlow, McMurray, and Cozens, 2004): above 0.90 indicates excellent reliability, between 0.70 and 0.90, indicates high reliability, from 0.50 to 0.70 indicates moderate reliability, and below 0.50 is low.

Finally, to estimate and visualize where responses were most reliable, we examined the correlation computed in each of the 128 channels using data from the 30 individual participants (topography of reliability).

Stability of response distribution across harmonics

We also tested the reliability between sessions in terms of the distribution of amplitude across the different harmonics of the individual face discrimination response (i.e., amplitude at 1.2, 2.4 Hz, etc.). The distribution of amplitude across harmonics in the frequency-domain is related to the morphology of the individual face discrimination response in the time domain (Retter & Rossion, 2016) and may hold additional individual information to estimate reliability. For example, two participants could have the same overall EEG amplitude, but the response could be distributed differently over different harmonics (see Figure 5A). This would indicate that the shape of their discrimination response differs. Hence, this analysis tests the stability of the shape of the response, or its qualitative aspects. We used a multivariate decoding approach to determine whether the pattern of amplitudes across harmonics allows distinguishing individual participant. For each participant and each session, we took the amplitudes at the significant harmonics of the face discrimination response (excluding the 5th harmonic corresponding to the 6 Hz general visual response) measured in 9 occipito-temporal electrodes in each hemisphere (in addition to the 5 channels included in the ROI, the neighbouring 4 electrodes were added to capture potential individual differences) resulting in one 108 features pattern per session. Each pattern was then normalized (z-scored) independently, to remove the effect of variations of amplitudes and isolate the contribution of the relative distribution of the neural response across harmonics. We used a simple Pearson correlation winner-take-all classifier using the first session as the training data set and the second session as the test data set. Participant decoding performance was taken as the averaged classification performance across participants. The

chance level was 3.3% and decoding performance was statistically evaluated using a permutation test where the decoding procedure was repeatedly performed (2000 times) after shuffling the participants' labels.

Between vs. within-session reliability

In an additional analysis, the between-session reliability was compared to split-half reliability within a session. Since the split-half reliability measures the internal consistency of the data, it can be applied as a benchmark to which we can compare the between-session reliability. In this way, we can evaluate whether or not the time between the two sessions (~2 months) has an effect on reliability. Again, we used a bootstrap approach. For within-session reliability, in each session, (1) 2 pairs of trials were randomly selected (e.g., pair 1: trials 1 and 3; pair 2: trials 2 and 4) per participant, (2) amplitude was averaged per pair and (3) Pearson correlation and ICC was computed using the 30 pairs (for 30 participants) of data points. Steps (1) to (3) were repeated 5000 times for each session (thus 10000 times for the 2 sessions) to determine the mean within-session correlation and a 95% confidence interval. In the between-session correlation analyses, the same bootstrapping analysis was performed using at each bootstrap only 2 randomly selected trials per session (out of 4 trials) for a fair comparison with within-session correlation.

Reliability as a function of the number of sequences

Lastly, we used a modelling approach to determine the minimal number of stimulation sequences needed to achieve optimal reliability. For this analysis, we used data from the optimal ROI (LOT or ROT) defined for each participant (see above). In separate runs of the analyses, we used either 1, 2, 3 or 4 (i.e., all sequences) sequences per participant and session. Using a bootstrap approach: (1) an equal number of sequences (i.e., 1–4) were randomly selected separately for each participant and each session, (2) the individual face discrimination amplitudes from the selected sequences were averaged, (3) the ICC across session was then computed (using the 30 participants) and the coefficient was stored. Steps (1) to (3) were repeated 5000 times to determine the mean correlation obtained with a fixed number of randomly selected sequences (1–4 sequences) and a 95% confidence interval. Then we modelled the relationship between the number of sequences and between-session reliability with a power function ($y = a * x^b + c$) where y is reliability and x is the number of sequences averaged. Similarly to this analysis, we estimated how the signal to noise ratio (SNR) is affected by the number of sequences. SNR was calculated as a Z-score (i.e., subtracting the mean amplitude in the 20 frequency bins surrounding the signal and dividing by the standard deviation across surrounding bins), thus taking into account the variability across the frequency bins. Since the within- and between-session reliability did not differ (see results), we combined the sequences of the two sessions to increase the number of available sequences (i.e., 8 sequences). Averaging across multiple trials, in our case 60 sec sequences, increases the SNR by reducing the contamination of voltage fluctuations unrelated to the stimulus presentation.

Base rate response

As a final step, the reliability of the general visual response was also assessed by estimating the Pearson and ICC coefficients, and we tested whether using this response to normalize the individual face discrimination response could improve the reliability of the latter.

Results

Stability of EEG amplitude to individual face discrimination across sessions

On grand-averaged data in the frequency domain, we observed a large response at 6 Hz (and harmonics), peaking on medial occipital sites (Figure 2) in both sessions. This general response reflects the brain's synchronization to the visual stimuli in general, i.e., a mixture of low-level and high-level processes in this paradigm (Liu-Shuang et al., 2014). It is not the main focus of the study, although its analysis is reported below and used in normalization procedures.

The 1.2 Hz response, reflecting individual discrimination of faces, was clear on grand-averaged data, and characterized by its 6 significant harmonics (1.2 Hz to 8.4 Hz, excluding the 6 Hz corresponding to the general response, Figure 2A). This response was maximal over occipito-temporal sites, with a right hemispheric advantage (Figure 2), replicating

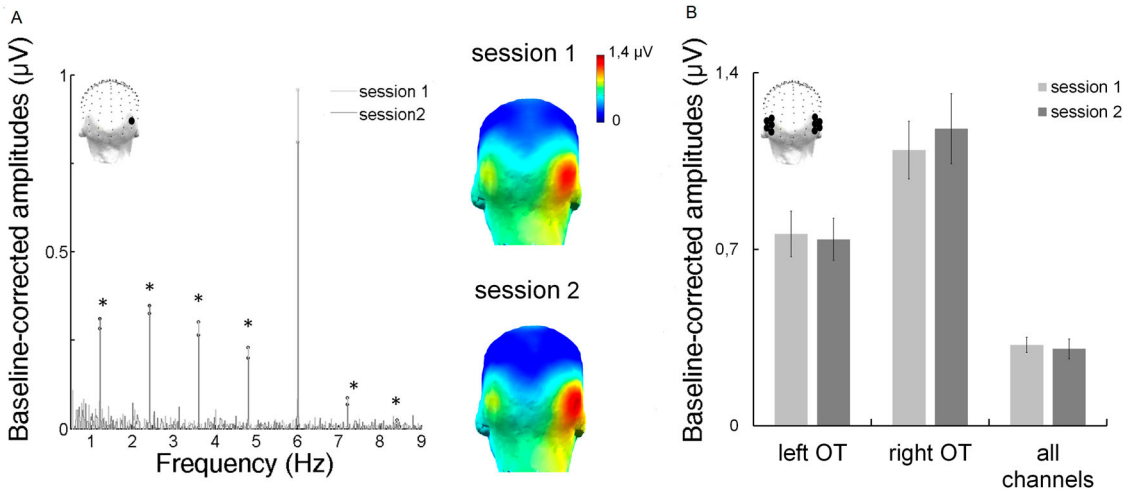

Figure 2. Facial identity discrimination response. A) Baseline-corrected amplitude spectra for the two sessions over right occipito-temporal electrode P10 highlighted with black circle on the blank headplot. Topographical maps show the baseline-corrected amplitudes summed over the significant harmonics of the facial identity discrimination response (until 8.4 Hz, marked with *). B) Summed baseline-corrected amplitudes for the discrimination response over left and right OT ROIs (channels are marked with black on the blank headplot) and averaged across all 128 channels.

all previous studies with this paradigm (Liu-Shuang et al., 2014; Dzhelyova & Rossion, 2014a, 2014b; Xu et al., 2017, 32 channels). It appeared virtually identical between the two sessions (Figure 2A).

The sum of the 1.2 Hz harmonics provides a quantification of the individual face discrimination response in each individual brain. This response was significant in all participants on a least 4 electrodes in session 1 at a conservative statistical threshold of $z > 3.1$, $p < 0.001$. In session 2, all participants also showed a significant response in at least 4 channels at this threshold ($Z = 3.1$, $p < 0.001$), except for one participant (P#18) who showed a focal response with only the left occipito-temporal PO9 channel reaching significance ($Z = 2.70$, $p < 0.0035$) and another participant (P#14) who had 8 significant electrodes but only at a lower threshold ($Z > 1.65$, $p < 0.05$ one-tailed).

The high reliability of the response at the individual level was supported by high correlations of the individual face discrimination response across individual participants ($N = 30$, $r = 0.79$; $p < 0.0001$, ICC = 0.87, 95% CI = [0.73 0.94]).

Stability of EEG spatial distribution to individual face discrimination across sessions

A stable individual face discrimination response was also observed in terms of the scalp distribution, both on grandaveraged group data (Figure 2) and individual data (Figure 3). On grandaveraged data, the correlation across scalp electrodes (i.e., 128 data points per session), reached 0.96 ($p < 0.0001$). In addition, intraclass and Pearson correlations computed at each of the 128 channels (Figure 4A, top row) revealed significant test-retest reliability mostly over occipito-temporal regions, particularly in the right hemisphere (Pearson correlations: range across right ROI electrodes: $r = 0.64$–0.79, $ps < 0.00005$, Figure 4A for ICC values, bottom row) where the maximal individual discrimination response was observed (Figure 2A) When computing reliability of scalp topographies at the individual level, the correlation and ICC dropped (averaged across the 30 participants, $r = 0.61$, ICC = 0.75, CI = [0.74 0.77]), most likely due to many channels not contributing to the discrimination response. Cronbach's alpha also showed high reliability ($\alpha = 0.76$) when considering all 128 channels and increased further ($\alpha = 0.82$) when considering only the 10 channels within the region of interest where most of the response was observed, thus eliminating random noise fluctuation introduced by the remaining channels.

The majority of the participants ($N = 23$) had maximal individual face discrimination response over the right hemisphere, with the exception of 7 individuals showing the largest response in the left hemisphere (P#1, P#3, P#16, P#18, P319, P#28, P#30; see Figure 3). Considering only the difference between the left and right hemisphere, all participants showed consistent hemispheric dominance across sessions except for 4

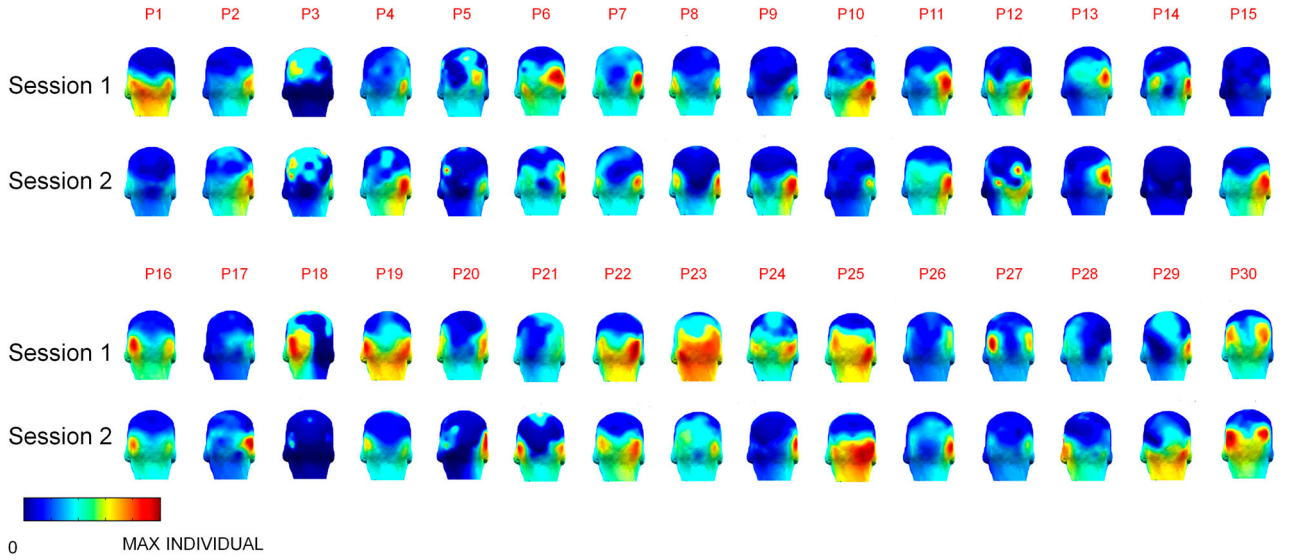

Figure 3. Individual topographies of the face discrimination response at the test-retest sessions. Summed (up to 8.4 Hz, excluding the 6 Hz response) baseline-corrected amplitudes are adjusted according to the individual's maximal value.

individuals (P#3, P#14, P#21, P#27). Yet, even for these individuals, the channels with maximal activation for the individual face discrimination response were located within the same hemisphere. Only P#27 showed a shift in the hemispheric dominance across the sessions, although a consistent activation was observed over the right hemisphere for both sessions.

Patterns of harmonics

In addition to the quantitative stability of the individual discrimination response, we investigated the between-session reliability of the pattern of individual face discrimination harmonics with a decoding approach, thus providing a measure of qualitative reliability. When using the 6 significant harmonics of 1.2 Hz (1.2–8.4 Hz, excluding 6 Hz) in the first session as training set, classification performance was at 30%, which was highly significantly above chance level (chance level: 3.3%; 99% confidence interval for chance: 13.3%, $p < 0.0005$, 1-tailed permutation test). Classification performance slightly increased further when using only the first 4 harmonics of 1.2 Hz, which are the most robust (Figure 2A): 33.3% (99% CI for chance: 13.3%, $p < 0.0005$).

Internal consistency (split-half reliability)

In addition, we investigated whether between-session reliability is limited due to the time separation between the 2 sessions (~ 2 months) or due to limited internal consistency. To achieve this we compared between-session reliability to within-session split-half reliability (i.e., internal consistency). Here we quantified between-sessions reliability using only 2 sequences (out of 4) per session to match the number of sequences used in the split-half within-session reliability. Impressively, the between sessions reliability ($r = 0.63$, 95% CI = [0.46 0.78], ICC = 0.78, 95% CI = [0.69 0.86] Figure 4D) was almost identical to the within-session reliability ($r = 0.65$, 95% CI = [0.53 0.78], ICC = 0.76, 95% CI = [0.63 0.87]), indicating that between-session reliability is limited by internal consistency rather than by time separation.

Reliability as a function of the number of sequences

In the last analysis, we tested how the number of 60 s stimulation sequences performed by each participant affects between-session reliability. This was done by computing between-session reliability (ICC) using data (using the optimal LOT or ROT ROI for each subject) from either 1, 2, 3 or 4 (i.e., all) stimulation sequences (Figure 4C). We found that reliability increases with the number of sequences used. Modelling this relationship suggests that moderate reliability could be expected even with just 1 min (1 sequence) of recording. Increasing the number of sequences leads to high reliability and including 5 or more sequences would lead to excellent reliability (Figure

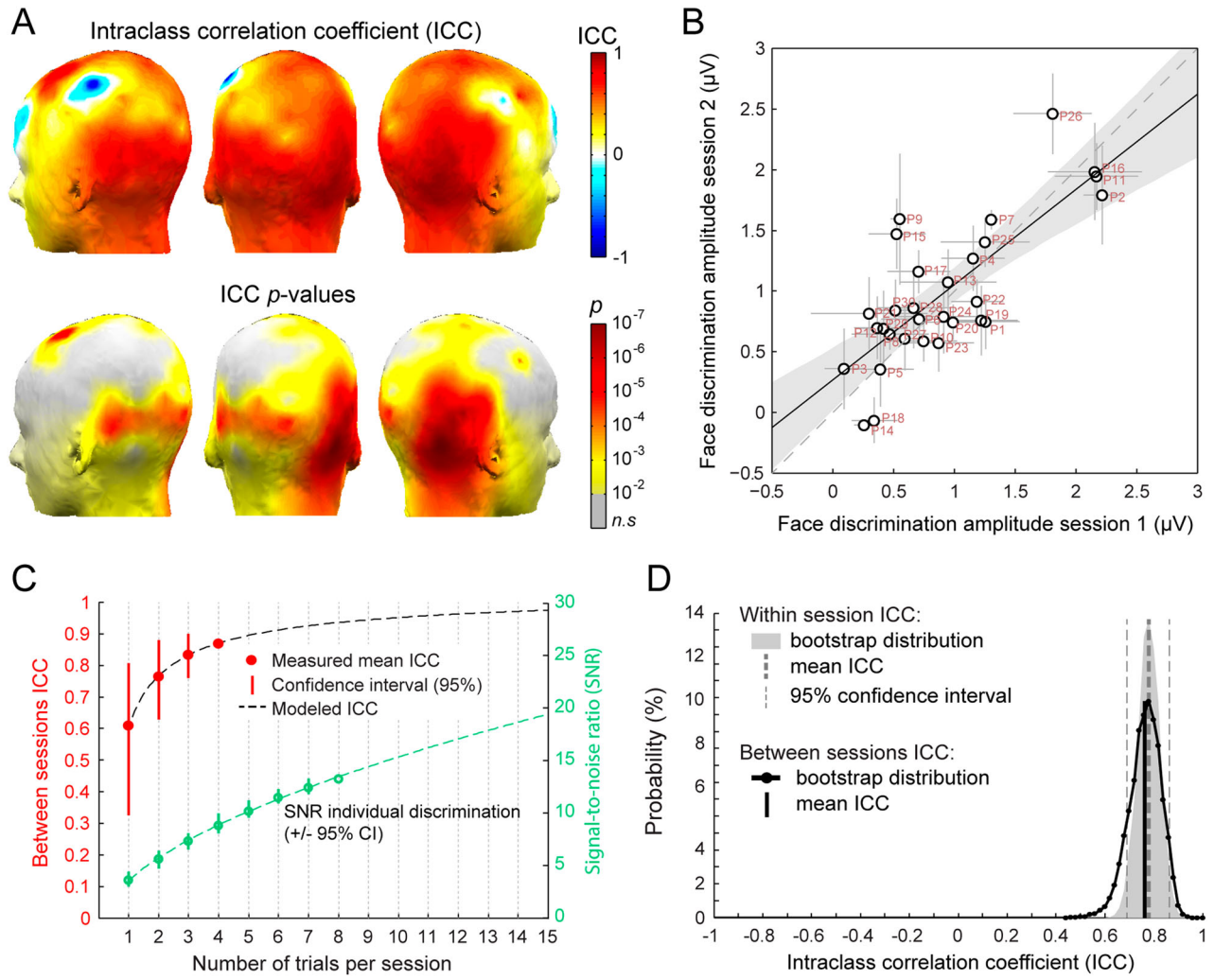

Figure 4. Reliability of the individual face discrimination response. A) Intraclass correlation coefficients across the two sessions (~ 2 months apart) for each 128 channels displayed as topographical maps (upper row) and the corresponding *p*-values (bottom row). Significant correlations ($p < 0.01$) are colour-coded; non-significant correlations are showed in grey. B) Scatter plot showing the individual face discrimination response (1.2 Hz) amplitude across the two sessions for all participants. Each data point represent the baseline-corrected amplitude summed across harmonics (up to 8.40 Hz, excluding the 6 Hz general response) for the optimal occipito-temporal ROI (LOT or ROT) per participant, and averaged across the 4 trials from each session. Error bars indicate SEM across the 4 trials of the same session. C) Model estimation of the relationship between number of trials, signal to noise ratio and the between session reliability. Red dots indicate measured intraclass correlations for 1, 2, 3 or 4 (all) trials as well as 95% confidence interval. Dotted line shows the fitted power-law function, which corresponding model, parameters, and goodness of fit (R^2) are indicated. Green dots represented the estimated signal to noise ratio D) Comparing within- and between-sessions reliability. Bootstrap distribution of correlations for within-session (light grey filled) and between-sessions (black line) reliability estimates. Correlations were computed using 2 randomly selected trials either from the same session (within) or different sessions (between) for each bootstrap. Thick vertical lines correspond to the distributions mean and thin dotted lines indicate the 95% confidence interval for within-session correlations.

4C). Similarly, increasing the number of presented sequences increased the signal to noise ratio.

Base rate response and normalization

The general visual response observed at 6 Hz and harmonics reflects a mixture of low- and high-level processes, which could therefore potentially fluctuate more between sessions than a more specific individual face discrimination response. Alternatively, since it reflects more basic processes, this response could also be more stable across individuals. Although this response was not of high interest in the current study, we also evaluated its reliability both for information and for using this response to normalize the individual face discrimination response.

The general visual response (i.e., quantified as the sum of 8 significant harmonics of 6 Hz, up to 48 Hz)

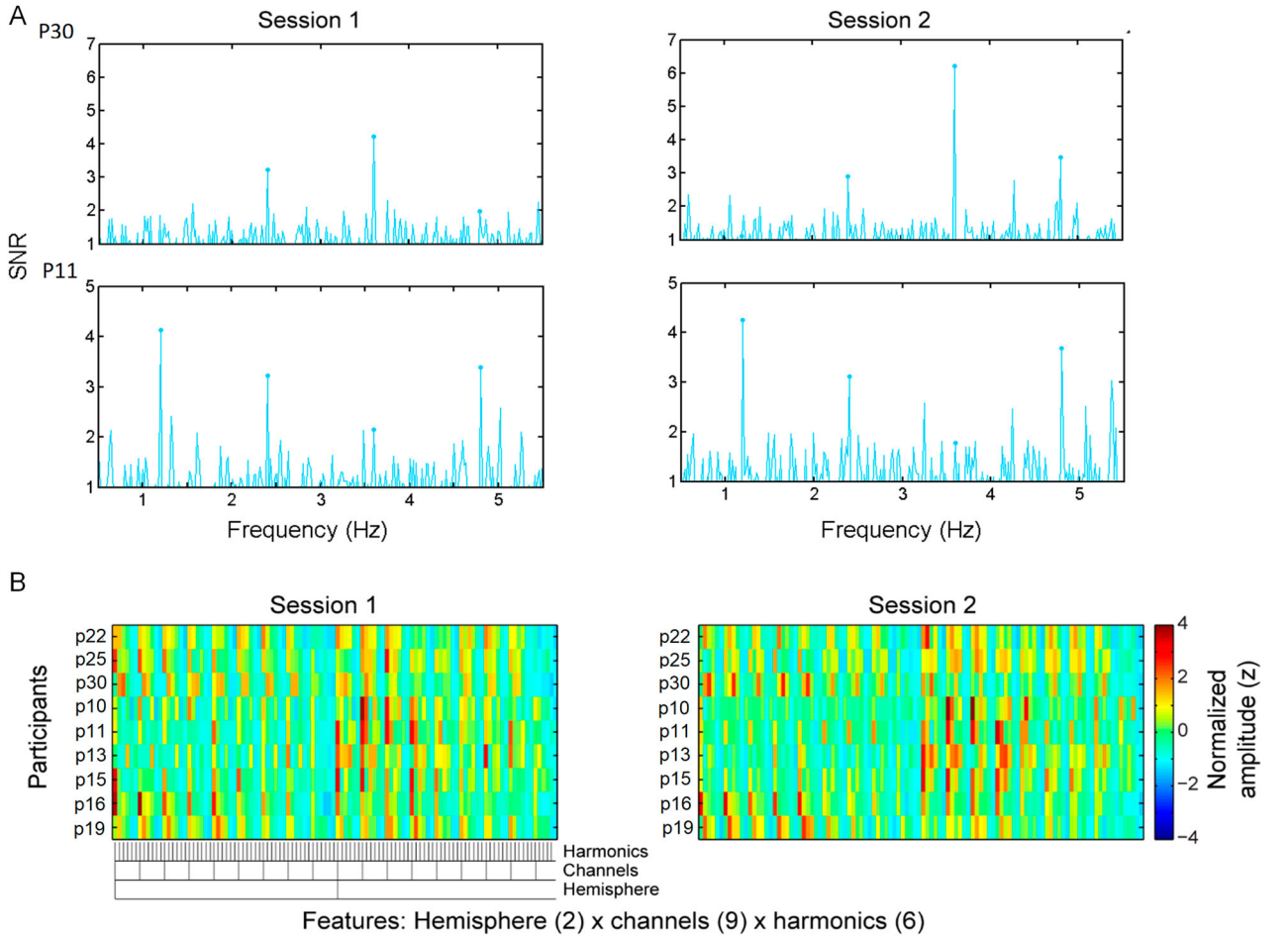

Figure 5. A) Example spectrum for two participants (P11 and P30) of the averaged 9 occipito-temporal channels over the optimal hemisphere. Harmonics are marked with a circle. B) Similarity in the pattern of harmonics across the test-retest sessions. Panels display the z-transformed amplitudes for each significant harmonic (1.2, 2.4, 3.6, 4.8, 7.2, 8.4 Hz) of the identity discrimination response across the 18 channels (9 per hemisphere- left hemisphere: TP7, P7, P9, PO7, PO9, PO11, O1, POI1, I1; right hemisphere: TP8, P8, P10, PO8, PO10, PO12, O2, POI2, I2) assessed in the classification analysis for 9 participants (P22, P25, P30, P10, P11,P13, P15, P16, P19).

was distributed over medial occipital regions. It peaked over channel Oz in both sessions: session 1 ($M = 2.56$ µV, $z = 216.09$); session 2 ($M = 2.28$ µV, $z = 126.88$). The reliability of the general response at the individual level was confirmed with a high correlation between the two sessions over the MO ROI ($r = 0.88$, $p < 0.0001$, ICC = 0.93, 95% CI = [0.86–0.97]).

In an additional analysis, we attempted to further isolate the specific individual face discrimination response by normalizing its amplitude with the amplitude of the general visual response (i.e., face discrimination response/sum of face discrimination response and general visual response). This procedure could potentially reduce the contribution of non-specific parameters affecting the overall EEG amplitude in each individual. To normalize the face discrimination, the general response over the medial region was used (i.e., normalization based on the respective maxima for each type of response). The reliability was very high ($r = 0.87$, $p < 0.0001$; ICC = 0.92, 95% CI = [0.82 0.96]), in fact reaching the highest reliability values for the individual face discrimination response.

Discussion

A highly reliable measure of individual face discrimination

At a group level, we found highly stable amplitudes and spatial distributions of a robust electrophysiological index of unfamiliar face discrimination across recordings performed at two months interval. Since the same unfamiliar face stimuli were used in the

two sessions (although their relative frequency of presentations and their order of appearance differed across sessions, see below), this suggests that any familiarity with these stimuli remaining from exposure at session 1 did not influence the measure at session 2, at least at the group level. This stability at the group level is already an interesting result compared to explicit behavioural tests, where test-retest effects could be observed because individual participants become familiar with the task or the specific stimulus set used. For example, 6 months apart test-retest effects have been reported at the CFMT, with substantial increases in accuracy from test to retest (mean = 76.9%, *SD* = 12.9% to mean = 83.2%, *SD* = 12.9%, Wilmer et al., 2010).

At the individual level, only four stimulation sequences, corresponding to four minutes of experiments, were sufficient to elicit a significant response in every individual participant at each recording session, even though this response was relatively low in amplitude for a few individuals. This observation confirms previous studies with the same paradigm, showing the high sensitivity of the FPVS-EEG approach to capture the individual face discrimination function in individual participants (Liu-Shuang et al., 2014; Liu-Shuang et al., 2016; Xu et al., 2017).

In both sessions, the EEG individual face discrimination response varied substantially across the 30 individuals tested in terms of amplitude (sum of harmonics in the EEG spectrum, range: 0.18–2.65 μV averaged across the two sessions; Figure 4), and to a lesser extent in scalp topography (Figure 3) and shape (i.e., the distribution of voltage amplitude across harmonics; Figure 5). Beyond a genuine variability in terms of the magnitude of the source of the neural response, presumably linked to the assessed brain function (i.e., individual face discrimination), several general factors could potentially account for the inter-individual EEG variability recorded on the scalp. For instance, even if the recorded measure directly reflects an index of individual face discrimination rather than the absolute EEG amplitude to the onset of a face, skull thickness and (mainly) the orientation of the functional gyri and sulci with respect to the location of the electrodes on the scalp should play an important role in the variable EEG amplitudes and scalp topographies recorded on the scalp. (Luck, 2005; Nunez & Srinivasan, 2006; Woodman, 2010). Presumably, these factors remain constant across sessions, and so should the measured function, i.e., individual face discrimination. Nevertheless, other factors potentially influencing EEG amplitude, such as fluctuations of attention, could well vary across sessions. In this context, the reliability of the response computed across individuals, which reached $r = 0.79$ in only 4 min of recordings, can be considered as particularly impressive. This reliability across individuals is reduced to about $r = 0.63$ when using only 2 stimulation sequences, i.e., 2 min of recording, indicating that in studies of interindividual differences, more than 2 sequences should be recorded. To be more precise on that matter, our modelization based on the reliability of the response for 1, 2, 3 or 4 sequences show that a total of 5–6 stimulation sequences (i.e., 5–6 min of recording) would suffice to reach (between sessions) reliability values of 0.95 and above (Figure 4C) after which adding additional trials did not substantially change the reliability of the response.

Given the lack of test-retest reports of measures of individual unfamiliar face discrimination with a relatively long interval between the two measures, comparing these reliability values to other measures of this function is difficult. However, here, quite impressively, the between-session reliability computed on two recorded sequences is virtually identical to the within-session reliability (Figure 4D). This indicates that the between sessions reliability is mostly influenced by within session noise rather than by changes occurring across sessions. Interestingly, although very little data is available on this issue, this does not appear to be the case for explicit behavioural tests. For instance, the split-half reliability of the CFMT is very high (0.83–0.89; Wilmer et al., 2010), but the reliability over a 6 months period drops to 0.70 (Wilmer et al., 2010) while it is 0.79 for the Caledonian face test after 30 min delay (Logan, Wilkinson, Wilson, Gordon, & Loffler, 2016) and 0.67 for the Kent face matching test, done 7.2 days apart (Fysh & Bindemann, 2018). Split reliability values are available for the Glasgow face matching test ($r = 0.81$; Burton, White, & McNeill, 2010) and the BFRT-c (with too few items) is much lower ($r = 0.61$), but the RT data is more reliable ($r = 0.88$; Rossion & Michel, 2018). However, to our knowledge, no data have been provided about the test-retest correlation with long delays. Moreover, while comparably high test-retest reliability of ERP components such as the N170 has been observed

(Cassidy et al., 2012, Huffmeijer Bakermans-Kranenburg, Alink, & van IJzendoorn, 2014), these measures do not reflect an individual face discrimination response, as can be observed for instance when immediately repeating individual faces on the N170 component (e.g., Heisz, Watter, & Shedden, 2006; Jacques, d'Arripe, & Rossion, 2007; Caharel, d'Arripe, Ramon, Jacques, & Rossion, 2009) or the N250 component (Schweinberger, Pickering, Jentzsch, Burton, & Kaufmann, 2002). To our knowledge, test-retest reliability coefficients of these other neurophysiological individual face discrimination effects have not been reported.

The stability of EEG amplitude indexing individual face discrimination is also remarkable if one considers that, for a given participant, the 4 base faces most likely differ between the two sessions, since they are picked randomly in a large set of stimuli ($n = 25$) for each stimulation sequence. Moreover, the specific individual face discriminations, their order of appearance and the specific size variations within the sequences, also differ completely between session 1 and 2. While even higher reliability measures could potentially be obtained by presenting the exact same sequences twice (and using the exact same stimulation sequences for each participant), the current procedure is preferable since it identifies a highly sensitive and reliable response independent of specific stimulus discriminations. Presumably, this high stability despite the inter-sessions variations in stimulation condition is due to a large number of individual face discriminations performed across 4 sequences (i.e., 72 discriminations, or "oddballs", in each sequence, for a total of 288 discriminations considered in the measure).

Besides the individual face discrimination response, we also found that the general visual response (i.e., synchronization to every stimulus, regardless of changes of identity) at 6 Hz and harmonics was even more stable across sessions than the individual face discrimination response. This could be because of a higher variability across individuals of this response, which reflects a mixture of low- and high-level processes. Using the amplitude of this basic response at its peak (i.e., medial occipital channels) to normalize the amplitude of the individual face discrimination response further increased reliability values for the latter. While being potentially informative, we should also be wary of this procedure however, since a reliability measure could be essentially or completely driven by variability at the base rate response.

Reliability in qualitative aspects of the response: scalp topography and harmonics

Higher reliability values were obtained when considering the response over the occipito-temporal cortex only rather than on all channels across the scalp. This is because, despite some degree of interindividual variability in scalp topography, the individual face discrimination response focuses on (right) occipito-temporal channels in most individuals (Figure 3), again as in previous studies (Dzhelyova & Rossion, 2014a, 2014b; Liu-Shuang et al., 2014; Xu et al., 2017; see also Alonso-Prieto et al., 2013; Rossion et al., 2012; Rossion & Boremanse, 2011). In this context, unless an individual's response is maximal at other electrode locations, including lower EEG responses at other locations on the scalp than the occipito-temporal cortex should not be recommended, since these responses are noisier and less stable across sessions. Hence, an a priori selection of channels of interest based on current observations (i.e., maximal response over the scalp on grand-averaged data) and knowledge derived from previous investigations (Alonso-Prieto et al., 2013; Dzhelyova & Rossion, 2014a, 2014b; Liu-Shuang et al., 2014; Rossion et al., 2012; Rossion & Boremanse, 2011; Xu et al., 2017) is important to maximize reliability. Note however that the right hemispheric dominance is a group level effect, but that several individuals have a larger response over left occipito-temporal channels (Figure 3), indicating that the response over both hemispheres should be considered in the measure.

At the group level, the scalp topography is stable between sessions, and this is also the case at the individual level, as illustrated in Figure 3. The shape of the individual discrimination response (i.e., the deviation from the 6 Hz response), characterized by a specific pattern of harmonics, also varies across individuals (Figure 5), and can be used in addition to amplitude variations to predict well above chance level which individual's response is recorded at session 2 based on session 1 data. Although it may be (even more) difficult to relate variations in scalp topographies and patterns of harmonics, to performance at individual face discrimination, this genuine (i.e., stable) interindividual variability could be useful in clinical

studies. For instance, a rapid drop of amplitude or a qualitative change in scalp topography/pattern of harmonics between two recording sessions could be used as a flag for emerging difficulties in individual face discrimination for instance, as in Alzheimer's disease (e.g., Lavallée et al., 2016), Autism Spectrum Disorder (ASD, Vettori et al., 2019) or other clinical conditions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by a PDR FNRS grant (T.0207.16) and INTER/FNRS/15/11015111/Face perception grant.

References

Adrian, E. D., & Matthews, B. H. (1934). The Berger rhythm: Potential changes from the occipital lobes in man. *Brain*, *57* (4), 355–385.

Alonso-Prieto, E., Van Belle, G., Liu-Shuang, J., Norcia, A. M., & Rossion, B. (2013). The 6 Hz fundamental stimulation frequency rate for individual face discrimination in the right occipito-temporal cortex. *Neuropsychologia*, *51*(13), 2863–2875.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. doi:10.1162/neco.1995.7.6.1129

Benton, A. L., & Van Allen, M. W. (1968). Impairment in facial recognition in patients with cerebral disease. *Cortex*, *4*(4), 344–IN1.

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching faces against the clock. *i-Perception*, *7*(5), 2041669516672219.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., … Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge face memory test and Cambridge face perception test. *Cognitive Neuropsychology*, 26(5), 423–455.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339–360.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*, 207–218.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286–291.

Busigny, T., & Rossion, B. (2010). Acquired prosopagnosia abolishes the face inversion effect. *Cortex*, *46*, 965–981.

Caharel, S., d'Arripe, O., Ramon, M., Jacques, C., & Rossion, B. (2009). Early adaptation to repeated unfamiliar faces across viewpoint changes in the right hemisphere: Evidence from the N170 ERP component. *Neuropsychologia*, *47*(3), 639–643.

Cassidy, S. M., Robertson, I. H., & O'Connell, R. G. (2012). Retest reliability of event-related potentials: Evidence from a variety of paradigms. *Psychophysiology*, *49*(5), 659–664.

Clayson, P. E., & Larson, M. J. (2013). Psychometric properties of conflict monitoring and conflict adaptation indices: Response time and conflict N 2 event-related potentials. *Psychophysiology*, *50*(12), 1209–1219.

De Renzi, E., & Spinnler, H. (1966). Visual recognition in patients with unilateral cerebral disease. *The Journal of Nervous and Mental Disease*, *142*(6), 515–525. doi:10.1097/00005053-196606000-00002

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

Dzhelyova, M., & Rossion, B. (2014b). Supra-additive contribution of shape and surface information to individual face discrimination as revealed by fast periodic visual stimulation. *Journal of Vision*, *14*(14), 15. doi:10.1167/14.14.15

Estudillo, A., & Bindemann, M. (2014). Generalization across view in face memory and face matching. *I-Perception*, *5*, 589–601. doi:10.1068/i0669

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science*, *4*(6), 170249.

Fysh, M. C., & Bindemann, M. (2018). The Kent face matching test. *British Journal of Psychology*, *109*(2), 219–231.

Galper, R. E. (1970). Recognition of faces in photographic negative. *Psychonomic Science*, *19*(4), 207–208.

Heisz, J. J., Watter, S., & Shedden, J. M. (2006). Progressive N170 habituation to unattended repeated faces. *Vision Research*, *46*, 47–56.

Herzmann, G., Danthiir, V., Schacht, A., Sommer, W., & Wilhelm, O. (2008). Toward a comprehensive test battery for face cognition: Assessment of the tasks. *Behavior Research Methods*, *40*(3), 840–857.

Hinton, P. R., Brownlow, C., McMurray, I., & Cozens, B. (2004). Using SPSS to analyse questionnaires: Reliability. In *SPSS explained* (pp. 356–366). London: Routledge.

Huffmeijer, R., Bakermans-Kranenburg, M. J., Alink, L. R., & van IJzendoorn, M. H. (2014). Reliability of event-related potentials: The influence of number of trials and electrodes. *Physiology & Behavior*, *130*, 13–22.

Jacques, C., d'Arripe, O., & Rossion, B. (2007). The time course of the face inversion effect during individual face discrimination. *Journal of Vision*, *7*(8), 3, 1–9.

Laguesse, R., Dormal, G., Biervoye, A., Kuefner, D., & Rossion, B. (2012). Extensive visual training in adulthood significantly reduces the face inversion effect. *Journal of Vision*, *12*(14), 1–13.

Lavallée, M. M., Gandini, D., Rouleau, I., Vallet, G. T., Joannette, M., Kergoat, M.-J., … Joubert, S. (2016). A qualitative impairment in face perception in Alzheimer's disease: Evidence

from a reduced face inversion effect. *Journal of Alzheimer's Disease*, *51*, 1225–1236.

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. doi:10.1016/j.neuropsychologia.2013.10.022

Liu-Shuang, J., Torfs, K., & Rossion, B. (2016). An objective electrophysiological marker of face individualisation impairment in acquired prosopagnosia with fast periodic visual stimulation. *Neuropsychologia*, *83*, 100–113.

Logan, A. J., Wilkinson, F., Wilson, H. R., Gordon, G. E., & Loffler, G. (2016). The Caledonian face test: A new test of face discrimination. *Vision Research*, *119*, 29–41.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Makeig, S., Bell, A. J., Jung, T. P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems* (pp. 145–151). Cambridge, MA: MIT Press.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876.

Munsters, N. M., van Ravenswaaij, H., van den Boomen, C., & Kemner, C. (2017). Test-retest reliability of infant event related potentials evoked by faces. *Neuropsychologia*, 126, 20–26.

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., & Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, *15*(6), 4.

Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of EEG* (2nd ed.). Oxford: Oxford University Press.

Regan, D. (1989). *Human brain electrophysiology: Evoked potentials and evoked magnetic fields in science and medicine*. New York: Elsevier.

Retter, T. L., & Rossion, B. (2016). Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream. *Neuropsychologia*, *91*, 9–28.

Rossion, B. (2014). Understanding individual face discrimination by means of fast periodic visual stimulation. *Experimental Brain Research*, *232*(6), 1599–1621.

Rossion, B. (2018). Prosopdysgnosia? What could it tell us about the neural organization of face and object recognition?. *Cognitive Neuropsychology*, *35*(1–2), 98–101.

Rossion, B., & Boremanse, A. (2011). Robust sensitivity to facial identity in the right human occipito-temporal cortex as revealed by steady-state visual-evoked potentials. *Journal of Vision*, *11*(2), 16–16.

Rossion, B., & Michel, C. (2018). Normative accuracy and response time data for the computerized Benton facial recognition test (BFRT-c). *Behavior Research Methods*, 50(6), 2442–2460.

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.

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

Schweinberger, S. R., Pickering, E. C., Jentzsch, I., Burton, A. M., & Kaufmann, J. M. (2002). Event-related brain potential evidence for a response of inferior temporal cortex to familiar face repetitions. *Cognitive Brain Research*, *14*, 398–409.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *British Journal of Psychology*, *75*(2), 221–242.

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology*, *54*(1), 123–138.

Vettori, S., Dzhelyova, M., Van der Donck, S., Jacques, C., Steyaert, J., Rossion, B., & Boets, B. (2019). Reduced neural sensitivity to rapid individual face discrimination in autism spectrum disorder. *NeuroImage: Clinical*, *21*, 101613.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., … Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*(11), 5238–5241.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, *72*(8), 2031–2046.

Xu, B., Liu-Shuang, J., Rossion, B., & Tanaka, J. (2017). Individual differences in face identity processing with fast periodic visual stimulation. *Journal of Cognitive Neuroscience*, *29*(8), 1368–1377.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145.

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences*, *22*(2), 100–110.