NeuroImage 58 (2011) 620–629

Contents lists available at ScienceDirect

NeuroImage

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

# Reliability of ERP and single-trial analyses☆

Carl M. Gaspar [a,*], Guillaume A. Rousselet [a], Cyril R. Pernet [b]

[a] Centre for Cognitive Neuroimaging (CCNi), Institute of Neuroscience and Psychology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
[b] Brain Research Imaging Centre, Division of Clinical Neurosciences, University of Edinburgh, Edinburgh, UK

ARTICLE INFO

*Article history:*
Received 24 November 2010
Revised 10 June 2011
Accepted 20 June 2011
Available online 25 June 2011

*Keywords:*
EEG
N170
Test–retest
Linear regression
Faces
Noise

ABSTRACT

A reliable measure is one we can trust in the long run. Thus, the reliability of measurements is as important as their validity. Here we investigated the reliability of brain electrical visual evoked responses to faces and noise textures. For the first time, we provide reliability measures for the full time course of event-related potentials (ERPs). Our analyses were also performed on a $R^2(t)$ metric that reflects results from single-trial analyses, therefore providing the first reliability analysis of ERP single-trial analyses. Results show that ERPs and $R^2(t)$ are highly reliable (cross-correlation ~0.9, lag ~4/6 ms, intra-class correlation ~0.9) but also idiosyncratic: ERPs and $R^2(t)$ are highly reproducible within subjects, who differ reliably from each other and the grand average across subjects. Consequently, grand averages, although highly reliable, can be misleading because they might not reflect the actual brain dynamic of any subjects.

## Introduction

This paper presents test–retest reliability results for event-related potentials (ERPs) to faces and noise textures. For the first time, we present results of test–retest reliability of the shape of the ERP time-course and the first test–retest reliability results for single-trial ERP analysis. The reliability of a scientific measurement places an upper limit on its validity. In neuroscience, reliability provides us with a magnitude of the noise that may mask brain signals, or be mistaken for them. Test–retest reliability has been reported for a variety of EEG measures: EEG power spectra (Gasser et al., 1985; Salinsky et al., 1991), variation of EEG power spectra across the spatial-frequency of checkerboards and gratings (SWEEP-VEP, Lauritzen et al., 2004), ERP peak responses to simple contrast patterns (Tello et al., 2010), and ERP peak modulations associated with both low-frequency stimuli (see Segalowitz and Barnes, 1993 for a review), and with performance feedback (Segalowitz et al., 2010). However, we are not aware of any published results on the test–retest reliability of ERP to objects and faces. This is a rather glaring omission because object perception, especially face perception, remains a dominant focus of ERP research on visual perception, giving rise to some of the most cited papers in that field.

### *ERP shapes*

Our paper focuses mainly on the temporal shape of ERP, which is not common. The bulk of ERP studies on object perception restricts their analysis to narrow time windows around peaks such as the P100 and the N170. Studies of general cognition also tend to focus on peaks, such as the P300. However, ERPs have significant correlational structure, suggesting that the entire shape of the ERP time-course may vary across either subject or experimental condition in interesting ways.

ERPs can be characterized not only by the latency and absolute amplitude of their peaks, but also by their entire shape. For example, Rousselet et al. (2009) find significant differences between young and old observers for the whole N170 component evoked by pink noise, even when ERPs were normalized within-subject across time and stimulus conditions: this suggests that there are reliable group differences, at least by age, in relative rather than absolute amplitude. Several other results showed that various ERP peak characteristics are correlated, both between- and within-subject. For instance, the latencies of the P1 and N170 to faces are correlated across variation in ages (a between-subject correlation - Kuefner et al., 2010). The N170 and P2 component amplitudes to partly phase-randomized faces are correlated within-subject (Rousselet et al, 2008a). The amplitude of the N170 and a subcomponent of the P300 (the late P3b) vary with working memory load in opposite ways during encoding of face identity, but vary with load in the same way during retrieval (Morgan et al., 2008). One alternative to peak-based analysis is to examine how ERP amplitude correlates with stimulus parameters at every time point (see Section 1.2, Single-trial ERP). Both the

☆ GRANT=ESRC grant RES-000-22-3209 supported GAR.
* Corresponding author.
*E-mail address:* Carl.Gaspar@glasgow.ac.uk (C.M. Gaspar).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.06.052

magnitude and the latency of these ERP-stimulus correlations are correlated with the amplitude of conventional ERP peaks (Philiastides et al., 2006; Schyns et al., 2007).

Variations in ERP shape also reflect interesting dynamics in information processing. Van Rijsbergen and Schyns (2009) have closely examined the temporal evolution of face emotion perception by reverse correlation between single-trial ERP and bubbled face-stimuli. They demonstrate that different time-frames, both prior to and after the N170, are associated with sensitivity to information from different parts of the face, and different spatial frequencies. Information sensitivity appears to be progressive, such that later processing supports visual representations that are a refinement of earlier representations. Philiastides and Sajda (2006) found, using a pattern classifier to measure the information content of single-trial ERP data during a face discrimination task, that various aspects of the information time course are consistent with the diffusion model (Laming, 1968; Ratcliff and McKoon, 2008). For example, a late surge in ERP information content occurs systematically later with lower quality in stimulus information, and the magnitude of this late surge is better correlated with subject accuracy than an earlier surge. Whether the time course of ERP reflects a refinement of visual representation, a general accumulation of visual information, or both, it is clear that the temporal shape of ERP can provide valuable clues about processing dynamics beyond what can be inferred from data restricted to ERP peaks and windows.

In this study, we measure the test–retest reliability of ERP shapes during face perception. To measure shape similarity we use the maximum of the temporally shifted correlations between two ERP waveforms (cross-correlation), in a temporal window from 0 to 500 ms. We also measure the temporal lag giving rise to the maximum correlation between two ERP waveforms.

### *Single-trial ERP*

The overall shape of ERP is difficult to interpret. As an alternative to mean ERP to faces, single-trial ERP approaches relying on linear modeling have been developed (Kiebel and Friston, 2004; Pernet et al., 2011 — see Rousselet et al., 2010 for an application).

In a previous experiment (Rousselet et al., 2008b), EEG was recorded while observers discriminated 2 faces. Because the phase of an image's Fourier components carries much of the information about object identity (Gaspar and Rousselet, 2009; Oppenheim and Lim, 1981; Piotrowski and Campbell, 1982), we degraded the natural appearance of the face stimuli by manipulating image phase along a single continuum from original phase —(maximal coherence) to completely random phase (no coherence). The use of this parametric design allows the expression of EEG in terms of sensitivity to stimulus information. For each subject, we analyze ERP using a single-trial general linear model. The linear model included the face identity (one of two), the phase coherence, the skewness and kurtosis of the image pixel distributions and their interactions with the phase as factors. The analysis of the different regression coefficients revealed that the N170 was sensitive to phase and kurtosis information, probably reflecting sensitivity to local edges. The time course of the coefficient of determination, $R^2(t)$, can also be used as a result of this analysis. The $R^2(t)$ function allows us to identify variations in neural activity across trials that are statistically associated with changes to visual information as modeled by the design matrix (i.e. the factors). Strong associations at certain time-points imply that the visual system activity is significantly modulated by image characteristics. For each subject, we can thus obtain a single $R^2(t)$ function, as an alternative to the traditional mean ERP measured in response to noise-free stimuli. In a follow up experiment, Rousselet et al. (2010) demonstrated systematic changes in $R^2(t)$ shape with age. For example, the latencies for 50% of the cumulative $R^2(t)$ could be predicted by subjects' age. This result suggests systematic variations in processing speed (information accumulation) with age, and also demonstrates that the shape of $R^2(t)$ as measured using a simple cumulative statistic, can provide interesting data. As a first step toward establishing the reliability of single-trial analysis, we measure the test–retest reliability of $R^2(t)$ shape during face perception. To measure shape similarity we use the maximum of the temporally shifted correlations between two $R^2(t)$ (cross-correlation) in a temporal window from 0 to 500 ms. We also measured the temporal lag giving rise to the maximum correlation between two $R^2(t)$.

### *Summary*

Our analyses aimed at answering two questions: 1st, are ERPs and $R^2$ idiosyncratic? and 2nd, is the within subject variance lower than between subject one?

To answer the 1st question, our analysis of reliability tests the hypothesis that ERPs and $R^2(t)$ functions are more similar when they are measured from the same person on different days, compared to when they are taken from different persons. Confirmation of this hypothesis raises the possibility that an individual's ERP or $R^2(t)$ function can be taken as their temporal signature of processing. We make no assumption about when the relevant idiosyncrasies might occur, or what form they may take. Therefore, our main measure of signal similarity (cross correlation) takes into account a large temporal window, 500 ms from the onset of the stimulus. There is also the possibility that some individuals are more reliable than others. While the behavioral equivalent is familiar to psychophysicists under the guise of internal noise (Burgess and Colborne, 1988; Gaspar et al., 2008), individual differences in the reliability of brain responses, as measured with EEG, have mostly been neglected. This is because traditional measures of reliability provide only a single value that reflects variation attributed to overall individual differences, relative to variation attributed to some experimental manipulation. In this paper, we take advantage of bootstrap statistics to provide a measure of reliability for each subject. To answer the 2nd question, we used a conventional statistic, i.e. intra-class correlation (ICC), to compare the relative reliability of different time frames, and also the relative utility of various ERP and regression metrics for predicting individual differences in other factors (beta swapping). Unlike the cross correlation analysis, our ICC-based results do not provide information about individual differences in shape reliability; instead they provide information about differences in the reliability of different time-frames, and single-valued statistics that can characterize ERP peaks or specific aspects of the $R^2(t)$ shape.

## Materials and methods

### *Participants*

Five University of Glasgow observers (GAR, ERN, MRL, KZA and DVD) took part in the experiment. DVD was paid £6 per hour. All five observers had normal vision and gave informed consent prior to involvement. Glasgow University Faculty of Information and Mathematical Sciences Ethics Committee approved the project.

### *Experimental design*

Testing was done between 8 a.m. and noon, and never after the subject had eaten lunch. Each day, during 5 days, subjects performed a one-interval, two alternative forced choice task between two faces. These faces were selected from a set of 10 faces, which are described in detail in previous publications (Gold et al., 1999; Husk et al., 2007; Rousselet et al., 2008a, 2009). Each subject saw either two male or two female faces. Subjects sat in a dimly-lit, sound-attenuated booth. As previously described, stimulus phase information was manipulated across trials. All subjects had 1500 trials per testing day. There was a greater range in phase coherence for 2 subjects (KZA and DVD). Subjects DVD and KZA saw 50 conditions along a noise-signal continuum, from 0 to 100% phase coherence, with increments of 2%. Subjects GAR, ERN and MRL, saw 50 conditions along a noise-signal continuum, from 0 to 47% phase coherence, with an average increment of 1.70%. Viewing distance was maintained with a chinrest, at 90 cm. The screen was 28°×21°, and the stimuli 9°×9°. Stimuli were presented at an average luminance of about 33 cd/m2, for about 53 ms. Subjects were given unlimited time to respond by pressing 1 or 2 on the numerical pad of the keyboard to indicate which face had been displayed. We told subjects to emphasize response accuracy, not speed. Each observer completed 5 days of testing. On each day of testing, there were 10 blocks of 150 trials: 1500 trials in total, with 30 trials per level of phase coherence. Within each block, there was an equal number of repetitions of each face and each phase coherence level. Each block was preceded by 6 practice trials that allowed subjects to learn the stimulus-key association. Behavior and EEG data from each day of each subject were analyzed separately.

### EEG recording and preprocessing

We used a 128-channel Biosemi Active Two EEG system (BioSemi, Amsterdam, Netherlands). We recorded from four additional electrodes - UltraFlat Active BioSemi electrodes — below and at the outer canthi of both eyes. Analog signal was digitized at 512 Hz and band-pass filtered online between 0.1 and 200 Hz. Electrode offsets were kept between +/− 20 μV. Practice trials were excluded. EEG was bandpass filtered between 1 and 40 Hz, following our observation that evoked responses to faces are contained within a narrow 5–15 Hz band (Rousselet et al., 2007). EEG was organized into epochs from 300 ms before stimulus onset, to 1200 ms after. Mean baseline activity, measured 300 ms before stimulus onset, was subtracted from every time point, and we used an average reference. Noisy cortical electrodes (non-ocular) were visually identified and removed; remaining electrodes averaged 117 [range 102–128]. An ICA was performed (Makeig et al., 2004), as implemented in the runica EEGLAB function (Delorme et al., 2007; Delorme and Makeig, 2004). We removed ICA corresponding to blink activity, identified by visual inspection of their scalp topographies, time courses and activity spectra. Then EEG was reorganized into shorter epochs from 300 ms before stimulus onset, to 500 ms after stimulus onset. Baseline activity, measured in the 300 ms before stimulus onset was subtracted again. Remaining trials with artifacts were rejected. Artifactual trials were defined as trials with EEG greater than ±100 μV, or a linear trend>75 μV/epoch and with a regression $R^2$>0.30. All remaining trials were included in the analyses, whether they were associated with correct or incorrect behavioral responses. The average number of trials per day was 1465, minimum 1266, maximum 1500.

### ERP measurement

We calculated mean face and mean noise ERPs from electrodes on the scalp's posterior. For each observer on each day (25 data sets), we selected the electrode exhibiting the largest N170 component in response to faces; the same electrode was selected for noise ERPs. Trial and error revealed that electrodes with the largest N170 component showed the largest difference between the maximum face-ERP voltage between 50 and 150 ms, and the minimum voltage between 100 and 220 ms. Reliability analyses were then performed on the mean ERPs for these 25 electrodes (1 electrode per day per subject).

Because $R^2$(t) functions are naturally based on a large number of trials (here 1465 trials on average), we also used a high number of trials to create ERPs. While we are not directly comparing the reliability of ERP to $R^2$(t), we did not want to obtain relatively weak ERP reliability based only on small sample size. Therefore, our ERPs are based on the largest number of trials (240 trials) that is consistent with a published ERP study that uses the same behavioral task (Rousselet et al., 2008b), or similar stimuli (Rousselet et al., 2004). Here, we obtained two separate ERPs for each person and each day, leading to a total of 25 face ERPs and 25 noise ERPs. Face ERPs were obtained by averaging the 240 trials with the highest coherence level; noise ERPs were obtained by averaging the 240 trials with the lowest coherence levels.

Different ranges of image phase coherence were shown to the first 3 observers (GAR, ERN and MRL), and to the last 2 observers (DVD and KZA). The 240 trials with the highest-coherence trials formed the ERPs to faces, Eface(t), (around 47% for the first three subjects, and 100% for the last two subjects). The use of 240 face trials for the ERP resulted in coherence levels giving predicted accuracies of at least 95% for all observers. To ensure that the sampled trials resulted in a dichotomy between chance and above 95% accuracy, we measured predicted accuracy, based on psychometric function fit, across all culled trials, and, for each observer, took the mean across trial and day. The 240 trials with the lowest coherence formed the ERPs to noise, Enoise(t). For the lowest-coherence trials, mean predicted accuracy was exactly 50% for each of the 5 observers. For the highest coherence trials, mean predicted accuracy was 98%, 100%, 96%, 100% and 100%, for observers ERN, GAR, MRL, KZA and DVD, respectively. Thus, despite using different phase coherence range, the face and noise ERPs reflected averages coming from similar performance ranges.

### Grand-average ERP

Ten additional grand-average ERPs of Eface(t) and Enoise(t) were obtained by averaging ERPs across subjects separately for each day of testing.

### $R^2(t)$ measurement

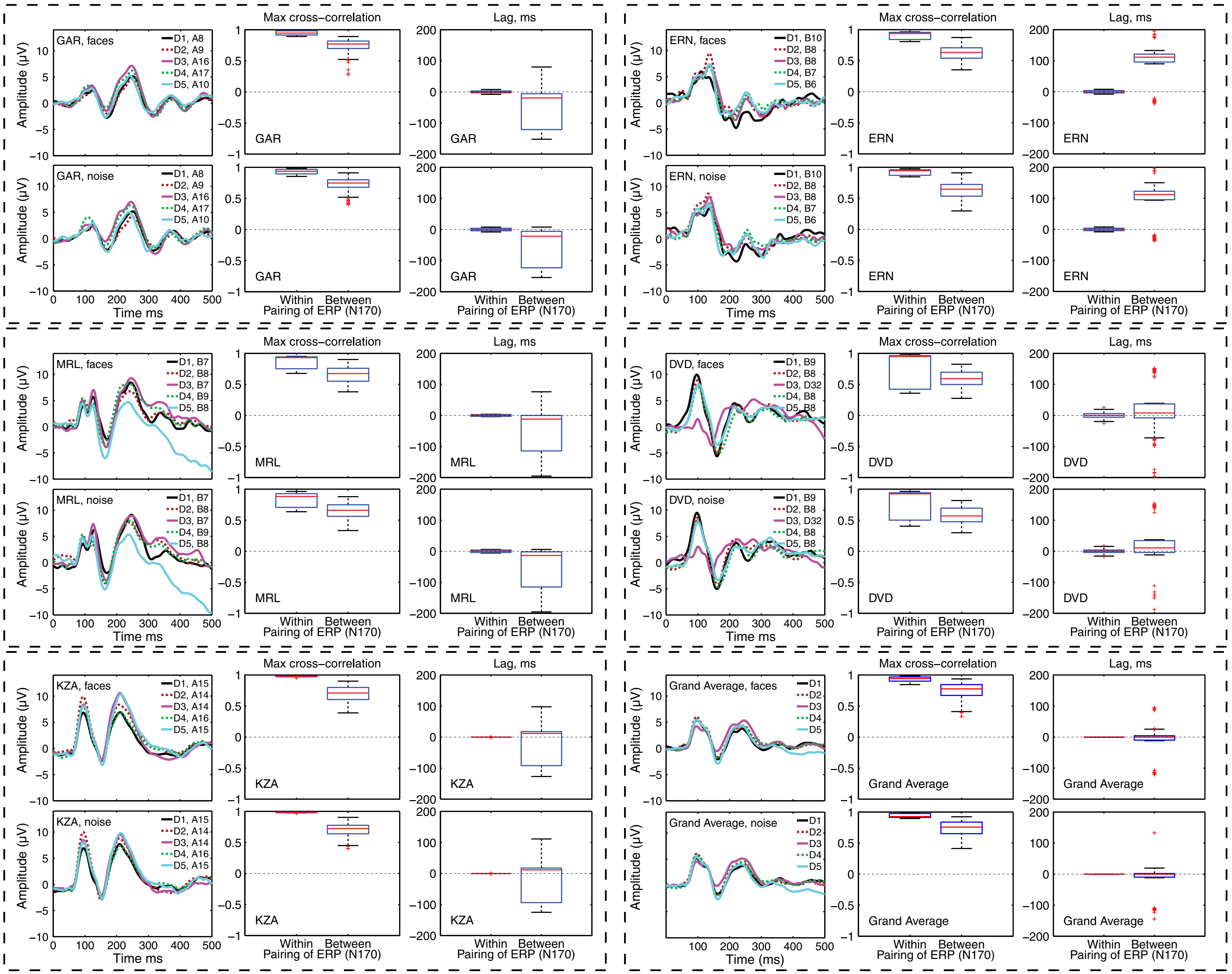
Data were analyzed using the LIMO EEG toolbox (Pernet et al., 2011 — https://gforge.dcn.ed.ac.uk/gf/project/limo_eeg/). Using a general linear model (GLM) approach, single-trial ERP amplitudes (independently at each time point and each electrode) were modeled such as:

$$ERP = \beta 0 + \beta 1F1 + \beta 2F2 + \beta 3\varphi + e. \quad (1)$$

The regressor for stimulus identity, i.e. face 1 (F1) vs. face 2 (F2), was a categorical factor. Global phase coherence (φ) was a continuous regressor. Global phase coherence is our image noise manipulation. β0 is the constant term and e the error.

For each subject on each day, we selected the electrode that showed the largest R2 over the entire data space, i.e. we searched for

**Fig. 1.** Plots of ERPs and ERP reliabilities for each subject, grouped by dashed boxes. The bottom-right group of plots depict results for the grand-average, modeled as a subject in its own right. The first row of 3 plots for each subject (dashed box) depicts results for face stimuli, while the second row of plots shows results for noise stimuli. The first, second and third columns of plots for each subject depict ERPs, maximum cross-correlation, and correlation lag, respectively. ERPs for each day of testing are shown using a different combination of color and line pattern. Maximum cross-correlation and correlation lag are measures of reliability. For both reliability measures, and for each subject, within- and between-subject distributions were compared using either a test of medians (for cross-correlations) or a test of MAD (for lags). For ERPs to face stimuli, failure to reject the null hypothesis occurred for only 1 of the 10 tests: DVD cross-correlations, p = 0.32. All other p values were equal to 0. Additionally, the null hypothesis was rejected for the grand-average 'subject'. For ERPs to noise stimuli, failure to reject the null hypothesis occurred for only 1 of the 10 tests: DVD cross-correlations, p = 0.26. All other p values were equal to 0. Additionally, the null hypothesis was rejected for the grand-average 'subject'.


GAR, faces
D1, A8
D2, A9
D3, A16
D4, A17
D5, A10
GAR, noise
Max cross–correlation
Lag, ms
GAR
Within
Between
Pairing of ERP (N170)
Amplitude (μV)
Time ms
ERN, faces
D1, B10
D2, B8
D3, B8
D4, B7
D5, B6
ERN, noise
ERN
MRL, faces
D1, B7
D2, B8
D3, B7
D4, B9
D5, B8
MRL, noise
MRL
DVD, faces
D1, B9
D2, B8
D3, D32
D4, B8
D5, B8
DVD, noise
DVD
KZA, faces
D1, A15
D2, A14
D3, A14
D4, A16
D5, A15
KZA, noise
KZA
Grand Average, faces
D1
D2
D3
D4
D5
Grand Average, noise
Grand Average
Time (ms)
0
100
200
300
400
500
−10
−5
0
5
10
−1
−0.5
0.5
1
−200
−100
200

the maximum R2 over electrodes and time frames and then focused our reliability analyses for the entire time course of that electrode. The signal at that particular electrode was most sensitive to the structure of the image as described by the design matrix and therefore constitutes the most likely candidate for reflecting the activity of cortical sources sensitive to image information. In general, $R^2(t)$ was the largest at posterior electrodes that also exhibited large responses to faces (Rousselet et al., 2008a). We obtained 25 different $R^2(t)$, one for each subject in each day.

### Reliability analyses

Our analyses addressed two questions: first, whether the shapes of ERPs and $R^2(t)$ are idiosyncratic, by measuring shape reliability (separately for each subject); and second, whether each time frame of the ERP or $R^2(t)$ is reliable in an absolute sense, by using ICC to measure intra-subject reliability separately for each time frame.

#### Cross-correlation

Cross-correlations were computed between all pairs of 25 face ERPs, all pairs of 25 noise ERPs and all pairs of 25 $R^2(t)$ (within- and between- subjects) and the signal agreement was based both on the maximum of the cross-correlation and its corresponding lag. The correlation is the Pearson product-moment correlation coefficient, ranging between +1, maximum agreement, and −1, maximum disagreement. The maximum cross-correlation is therefore the maximum value for the correlation coefficient obtained by temporally shifting the 2 signals with respect to each other. For example, two subjects may possess the exact same shape in $R^2(t)$, except for a 30 ms shift. In that case, the maximum cross-correlation will provide a high estimate of agreement (+1). The lag, measured in milliseconds, is the temporal shift required to obtain the maximum cross-correlation (e.g. 30 ms). For each subject, we obtained 6 separate distributions of 10 within-subject values ( 5!/(2!(5–2)!) ) for the maximum cross-correlation and the lag of face ERPs, noise ERPs, and $R^2(t)$. For each subject, we also obtained 6 distributions of 100 between-subject values (5 days × 4 other-subjects × 5 days).

#### Intraclass correlation

ICC was used across the whole ERP time courses for the same electrodes as the cross-correlation analyses. ICC describes how strongly observations from the same subjects resemble each other. While it can be viewed as a type of correlation, it operates on data structured by groups rather than data structured as paired observations. The ICC coefficient can thus be understood as a measure of discrimination between subjects (Bland and Altman, 1996). An ICC value of 1 indicates perfect reliability, i.e. that the signal amplitudes are identical from day to day within subjects while a value of 0 indicates no reliability. ICC temporal profiles were computed for each time points independently using the second class of ICC as defined by Shrout and Fleiss (1979).

$$ICC(2,1) = \frac{BMS-EMS}{BMS + (k-1)EMS + k(JMS-EMS)/n} \tag{2}$$

Eq. 2 estimates the correlation of the subject signal intensities between sessions, modeled by a two-way ANOVA, with random subject effects and fixed session effects. In this model, the total sum of squares is split into subject (BMS), session (JMS) and error (EMS) sums of squares. k is the number of repeated sessions, and n is the number of subjects. The ICC is a measure of absolute reliability instead of relative reliability, like Pearson correlation for example. This means that between-session variations in the mean of $R^2(t)$ (or ERP) across trials can lower reliability. ICC analysis is done separately at each time-frame and measures the absolute reliability of the signal at those time-frames. In contrast, the cross-correlation analysis is performed on the entire waveform and measures the reliability of shape, which is the relative value of $R^2$ (or ERP) across time.

#### β -swapping ($R^2(t)$ only)

For each subject on each day of testing, a specific set of model coeffcients (β) are estimated to fit the design matrix to the data. The modeled data give rise to $R^2(t)$ for that data. One can assess the goodness of model fit by measuring the mean-square error, or $MSE_{original}$. In this analysis, we apply the estimated β parameters for one data set (e.g. day 1) onto data from the same channel in another data set (e.g. day 2), and then measure how much the MSE increased using the ratio $MSE_{imposed}$ /$MSE_{original}$. Small increases in MSE suggest that estimated coefficients are interchangeable between the 2 data sets and therefore implies that data are highly reliable. Large increases suggest that the estimated coefficients are very different and therefore that data are not reliable. A subject who has a different set of model coefficients will exhibit significantly lower ratios of $MSE_{imposed}$ /$MSE_{original}$ for within- compared to between-subject comparisons. To illustrate how MSE may increase upon swapping model coefficients, consider the definition of MSE:

$$MSE = \sum_{i=1}^{f} \sum_{j=1}^{t} \frac{\left(\hat{y_{ij}} - y_{ij}\right)^2}{ft} \tag{3}$$

where f is the number of time-frames (257), and t is the number of trials (maximum of 1500), $y_{i,j}$ is the voltage at frame f and trial t. $y\hat{}_{ij}$ is the voltage predicted by the regression model. Note that $y\hat{}_{ij}$ varies across time because data are modeled separately for each time-frame. $y\hat{}_{ij}$ also varies across trials because the models are based on image phase-coherence, and face alternative, which vary across trials. The difference between $MSE_{original}$ and $MSE_{imposed}$ lies with $y\hat{}_{ij}$. For $MSE_{original}$, $y\hat{}_{ij}$ is based on the regression model that best fits $y_{ij}$ across trials j. In other words, $MSE_{original}$ is the conventional measure of mean-squared error. For $MSE_{imposed}$ , $y\hat{}_{ij}$ is measured by applying the beta- coefficients from one data set to another data set obtained from the same subject, the same time i and coherence-level for trial j, but on a different day (i.e. the design matrix is the same). Both $MSE_{original}$ and $MSE_{imposed}$ are measured using the same channel: the channel common to both data sets that gives the highest $R^2$ for the data originally used to obtain $y\hat{}_{ij}$ in $MSE_{imposed}$. Therefore, each $MSE_{imposed}$ has a corresponding $MSE_{original}$. Each ratio of ratio $MSE_{imposed}$ /$MSE_{original}$ has a minimum value of 1.

As for the cross-correlation analyses, the $MSE_{imposed}$ /$MSE_{original}$ ratio was computed for all possible pairs (within- vs. between- subjects). The distribution of within-subject comparisons comprised 20 pairs of $MSE_{original}$ and $MSE_{imposed}$ per subject: 5 days for the model used to obtain $y\hat{}_{ij}$, multiplied by 4 days of data inappropriate for those models. The distribution of between-subject comparisons comprised 100 pairs of $MSE_{original}$ and $MSE_{imposed}$ per subject: 20 sources (5 days × 4 subjects) for the regression model used to obtain $y\hat{}_{ij}$ in $MSE_{imposed}$, multiplied by 5 days of data inappropriate for those regression models.

#### Statistical analyses of reliability measures

We used percentile bootstrap tests to compare the medians of within- and between-subject distributions of cross-correlations and MSE ratios for each subject. Similarly, we used percentile bootstrap tests to compare the median absolute deviation from the median (MAD) of the distributions of cross-correlation lags. MAD is a robust measure of dispersion.

The percentile bootstrap test was computed as follow: (1) sample with replacement from the original distributions of within- and between-subject values; (2) calculate the estimator (median or MAD) of each resampled distribution; (3) compute and store the difference between the estimators of within- and between-subject values. In the case of cross-correlation values, the median of the between-subject distribution is subtracted from the median of the within-subject distribution, and in the case of MSE ratios, the subtraction is reversed. Steps one to three were performed one 1000 times. Two-tailed p-values were then computed as the proportion of median differences less than 0, or greater than 0, whichever is smaller. In all of our analyses, statistically significant differences are identified using a criterion p-value of 0.05.

## Results

### Behavior

The relationship between percent correct response and phase coherence was modeled by a Weibull function (Rousselet et al., 2009). From the fits, we obtained an estimate of subjects' 75% correct thresholds (Supplementary Fig. 1). Naturally, there is some variation in threshold across subjects. However, 75% thresholds are remarkably stable across days for each observer with an intra-class correlation coefficient of 0.98. This result suggests a low likelihood that within-subject variations in EEG are due to trivial factors such as lapses of attention.

### ERP

Eface(t) and Enoise(t) are shown in Fig. 1. For each subject, ERPs are very similar across days, except for MRL on day 5, and DVD on day 3.

### Cross-correlation analyses of ERPs

Reliability results for Eface(t) and for Enoise(t) are also shown in Fig. 1. For Eface(t), the mean of the median within-subject cross-correlation across subjects was 0.95 [0.78, -0.97] (inter-quartile limits are here and henceforth denoted by square brackets). The mean of median between-subject cross-correlations was 0.68 [0.58, 0.76]. The mean of the interquartile range of lag across subjects was 6 ms for within-subject comparisons, and 82 ms for between-subject comparisons. All subjects ($p = 0$) but DVD ($p = 0.32$) showed significantly higher within- than between-subjects median correlation distributions and all subjects showed significantly shorter within- than between-subjects lag dispersion distributions ($p = 0$). For Enoise(t), the mean of the median within-subject cross-correlation across subjects was 0.93 [0.79 to 0.96]. The mean of median between-subject cross-correlations was 0.67 [0.58, 0.75]. The mean of the interquartile range of lag across subjects was 6 ms for within-subject comparisons, and 81 ms for between-subject comparisons. As for Eface(t), all subjects ($p = 0$) but DVD ($p = 0.26$) showed significantly higher within- than between-subjects median correlation distributions and all subjects showed significantly shorter within- than between-subjects lag dispersion distributions ($p = 0$).

### Reliability of ERP grand averages

Reliability results for the grand-averages are shown in Fig. 1 (bottom right). For Eface(t), the median within-subject cross-correlation was 0.94 and the median between-subject cross-correlation was 0.78. The interquartile range of lag across subjects was 0 ms for within-subject, and 14 ms for between-subject comparisons. For Enoise(t), the median within-subject cross-correlation was 0.93 and the median between-subject cross-correlations was 0.76. The inter-quartile range of lag across subjects was 0 ms for within-subject and 12 ms for between-subject comparisons. For both measures of signal agreement, and for both types of stimuli (faces and noise), grand-averages were significantly different from actual subjects (all $p = 0$).

### Intraclass correlations of ERPs

ICC results for faces and noise are shown in Fig. 2. ICC values for both face and noise stimuli were very high (from .6 to .9), especially between 125 and 350 ms, which corresponds to the beginning of the N170 component and extends beyond the P2 component. Of interest, the ICC inter-quartile range was also very large in two narrow windows around ~125 ms and ~180 ms marking the beginning and the end of the N170 component.

### $R^2(t)$ analyses

$R^2(t)$ are shown in Fig. 3. For ease of visual comparison, each $R^2(t)$ is normalized by its maximum value.

### Cross-correlation analyses of $R^2(t)$

Reliability results for $R^2(t)$ are also shown in Fig. 3. The mean of the median within-subject cross-correlation across subjects was 0.94 [0.88, 0.97]. The mean of median between-subject cross-correlations was 0.70 [0.62, 0.79]. The mean of the interquartile range of lag across subjects was 4 ms for within-subject comparisons, and 74 ms for between-subject comparisons. All subjects ($p = 0$) but MRL ($p = 0.12$) showed significantly higher within- than between-subjects median $R^2(t)$ correlation distributions and all subjects showed significantly shorter within- than between-subjects lag dispersion distributions ($p = 0$).

### Intraclass correlations of $R^2(t)$

ICC results for $R^2(t)$ are shown in Fig. 2. Similar to ICC values for ERP, ICC values for $R^2(t)$ to both faces and noise can be very high (around 0.9). Unlike ICC for ERP, high values are maintained over a longer, uninterrupted time window, between from about 125 ms to 300 ms. Thus while ERPs were reliable from about the same period, they showed stronger variations than the $R^2(t)$ function especially at the beginning and the end of the N170 component (the ICC interquartile range for $R^2(t)$ at these latencies is much smaller). However, they also happen to be reliable for longer than $R^2(t)$, which can be explained by the fact the $R^2(t)$ only reflects the information of the design matrix.

### β-swapping

β-swapping results are shown in Fig. 4. For the analysis of maximum $R^2$ electrode, all subjects had MSE ratios significantly lower for within- compared to between-subject comparisons ($p = 0.018$ for MRL and $p = 0$ for all other subjects). The mean of the median within-subject MSE ratio across subjects was 1.029 [1.016, 1.077]. The mean of the median between-subject MSE ratio across subjects was 1.322

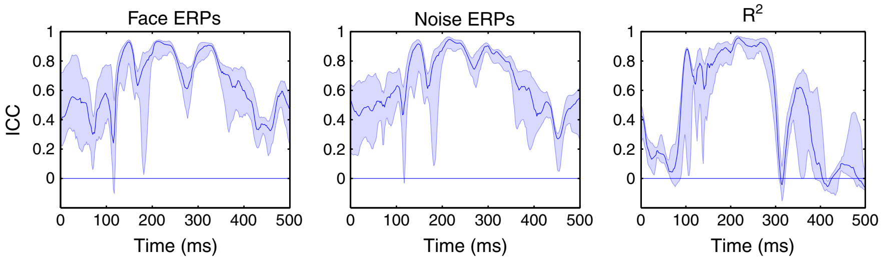


**Fig. 2.** Time course of test–retest ICC of ERPs to faces and noise, and of $R^2$ for max channel. Shaded area is the inter-quartile range obtained via bootstrap.

[1.249, 1.417]. For the analysis that includes all electrodes, all subjects had MSE ratios signicantly lower for within- compared to between-subject comparisons (p = 0). The mean of the median within-subject MSE ratio across subjects was 1.029 [1.02, 1.056]. The mean of the median between-subject MSE ratio across subjects was 1.137 [1.107, 1.179]. These results demonstrate that each subject's regression coefficients (estimated βs) are specific to their own EEG data, which is similar to say that ERP effects are reliable.

## Discussion

Our data suggest strong test–retest reliability of (i) the shape of mean ERPs to faces and to noise and (ii) of $R^2(t)$ functions. In both cases we observed stronger within than between subjects correlations and a stronger within than between subject variance (ICC analyses).

### *Relevance to past literature on ERP reliability*

Most of our main results are about the reliability of unusual features of ERP data sets: noise sensitivity (as measured using $R^2(t)$ functions) and shape of ERP waveforms. Therefore, it is difficult to relate our results to past studies that have examined the reliability of ERP peaks. Furthermore, our mean ERPs to faces were based on higher numbers of trials than usually used in comparable studies, and involved ICA to remove eye-blink components, an uncommon preprocessing step. Overall, our results provide measures of reliability

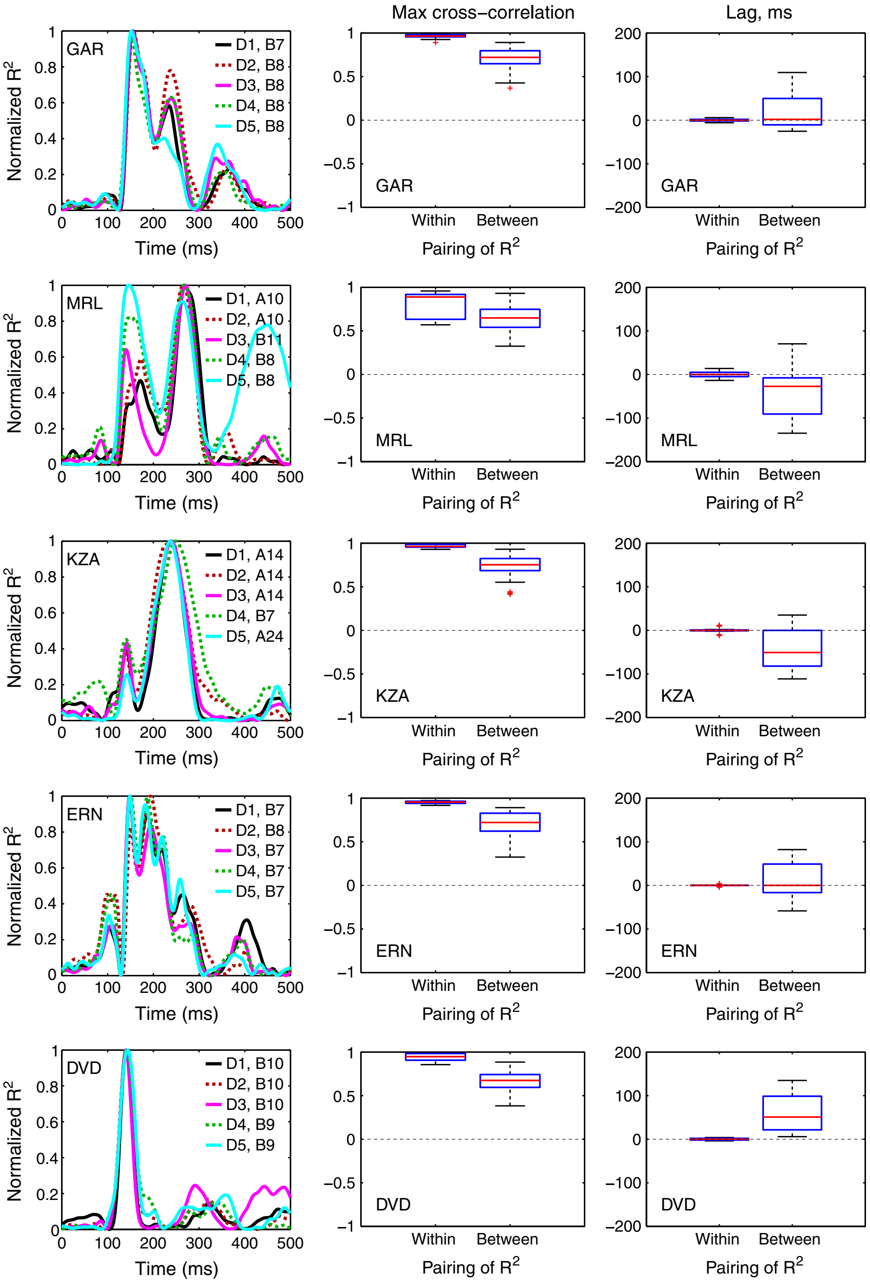


**Fig. 3.** Plots of normalized $R^2$ and their reliabilities for each subject, grouped by row. The first, second and third column of plots for each subject depicts $R^2$ functions, maximum cross-correlation, and correlation lag, respectively. $R^2$ functions for each day of testing are shown using a different combination of color and line pattern. Failure to reject the null hypothesis occurred for only 1 of the 10 subjects: MRL cross-correlations, p = 0.12. All other p values were equal to 0.

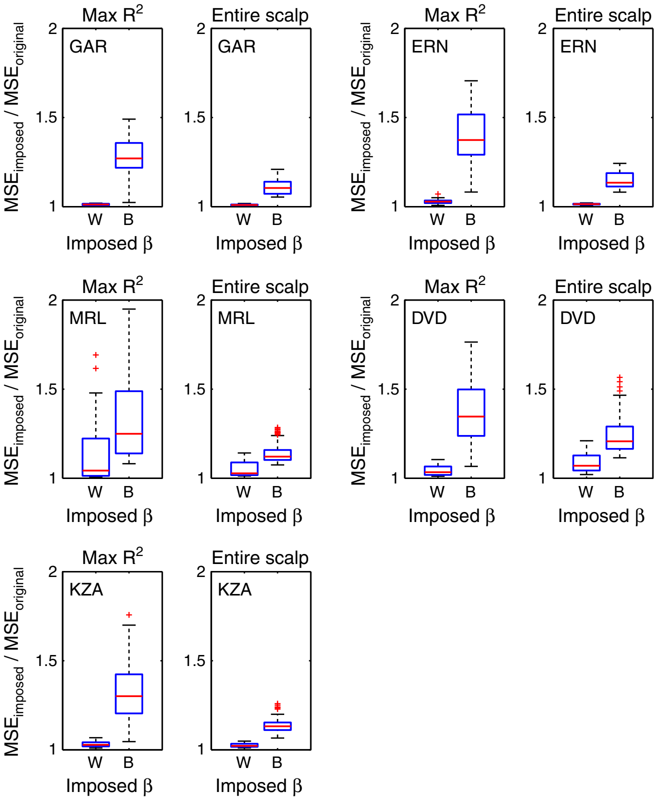


**Fig. 4.** Box plots of within- and between-subject MSE ratios are compared in separate plots for each subject, and each type of MSE: measured for the maximum $R^2$ channel, and for the entire scalp. All tests were statistically significant ($p = 0.018$ for MRL and $p = 0$ for all other subjects).

using novel measures and high-quality preprocessing, and pave the way for new kinds of ERP studies, rather than confirming more traditional methods (Rousselet and Pernet, 2011).

### *Temporal signatures of neural activity*

$R^2$ analyses revealed, as for ERP, a greater within than between subject reliability. It is important to remember that $R^2$ is a measure of model fit: the same $R^2$ values, for instance obtained at two different time frames, can correspond to very different values for the underlying model parameters. Therefore, the reliability of $R^2(t)$ and of the beta parameters are independent. That is why we measure the reliability of beta separately (Sections β-swapping ($R^2(t)$ only) and β-swapping ), and why it is not trivial that both $R^2(t)$ and beta were found to be reliable. In addition, even without modeling temporal dependencies, results indicate that entire temporal windows are reliable (see ICC analyses). This result could be partly explained by data filtering. However, filtering should also decrease between-subject differences, and ultimately stringent filtering might lead to equivalent values for within- and between-subject similarity of time-course. Therefore, the most likely conclusion is that individual differences in the shape of both mean ERPs and the $R^2(t)$ function can be viewed as a temporal signature for neural activity.

For illustrative purposes, we examine an extreme form of this view by treating mean ERP and $R^2(t)$ shape as a potential biometric marker. The standard test for the utility of a biometric marker is to assume a database of existing markers along with their identities, in this case either face ERPs, noise ERPs, or $R^2(t)$ functions, matched with subject identification. If we randomly sample markers without their identities, how accurate will identification be if we match the sampled marker to our database? We assume a procedure for matching that provides a single value that reflects degree of similarity, and chooses the identity with the highest value. This is the standard test applied to various biometric recognition systems, such as finger print and iris recognition systems, and the standard metric for accuracy is a non-parametric form of d-prime. D-prime is measured by comparing the distributions of match values for actual matches, and for actual mismatches. Our analysis here is purely illustrative because our matching procedure is a simple maximum cross-correlation, and is not optimized for biometric recognition; therefore, our results may be a gross underestimate of the true utility of mean ERP and $R^2(t)$ shape for biometric recognition. Furthermore, our sample size (5 subjects) is small. Nonetheless, these results can provide some indication about the potential for biometric recognition and, at the very least, provide some insight into whether mean ERP or $R^2(t)$ behave like temporal signatures for neural activity. Fig. 5 shows histograms that compare the distributions of within- (real matches) and between-subject (real mismatches) cross-correlations. Three plots and their respective d-prime scores, are shown, one each for face ERPs (=3.3), noise ERPs (=3.12), and $R^2(t)$ (=3.57). Basically, each plot collapses all the data from either Figs. 1, 3, or 4. For comparison, d-prime for biometric recognition using fingerprints is 6.8 (Jain, 2007), and varies between 7.3 and 14 for iris patterns, depending on whether the match is performed with the same camera or between cameras (Daugman, 2003). Considering that professional biometric systems are finely tuned to identify individuals for security reasons, and our EEG measures were not calibrated for such purposes, it is reassuring that our d-primes are only about half of d-primes obtained for these advanced systems. While we do not propose that our current measures can be used for biometric reasons, perhaps there is potential for this. In all cases, the ability to discriminate with relative ease each subject suggest that ERPs and $R^2(t)$ are unique to each individual.

### *Grand average ERPs can be misleading*

Published papers often show only the grand average ERP, and do not include plots of individual subjects. This can sometimes be a problem. As Kuefner et al. (2010) points out, sometimes conclusions are drawn from these plots that are not subjected to rigorous

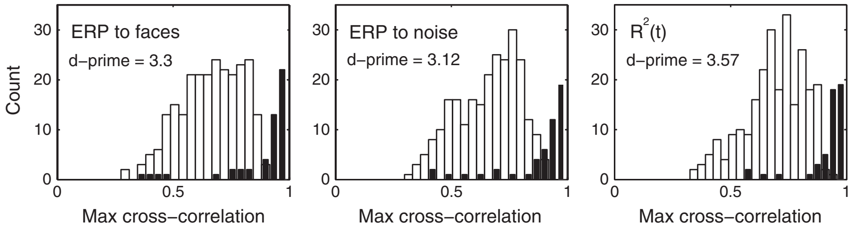


**Fig. 5.** The biometric sensitivity of ERPs to faces, ERPs to noise, and the $R^2$ function (in separate plots from left to right). Empty and solid bars for between- (identity-mismatches) and within-subject (identity matches) correlations, respectively. d-prime measures the separation between the 2 types of correlation.

statistical analysis. For example, Taylor et al. (2004) suggest that there are two distinct peaks after the N170, based on the appearance of the grand average ERP. But as Kuefner et al. (2010) suggests, such an average waveform could have been the product of individual variability in the timing of a single ERP peak. Results from Kuefner et al. (2010) are consistent with that conclusion. Our analysis of the grand averages to faces and noises also suggest that grand averages do not reflect well the ERP dynamics. Indeed, because ERPs are highly reliable within subjects (>0.90 with only 6 ms lag), the grand averages are also highly reliable. However, this 'within-subject' reliability also means that grand averages ERPs are significantly different from individual subjects' ERPs.

### Temporal and spatial dependence of reliability

Overall shape, measured over long time windows (up to 500 ms from stimulus onset), can provide valuable information about individual differences. However, our ICC results suggest that reliability may be improved by restricting analysis to a time window from 0 to 300 ms, especially for the single-trial analysis used to obtain $R^2(t)$ functions. Our β-swapping results show that all 5subjects were highly reliable. Nonetheless, the separation between within- and between-subjects MSE ratios might be further increased by performing analysis on a time window from 0 to 300 ms. We re-calculated results for MRL because he had the least separation between within- and between-subjects MSE ratios (Supplementary Fig. 2). Compared with MRL results on a time window from 0 to 500 ms, there is a marked increase in the separation between within- and between-subject distributions; specifically, there is a 62-percent reduction in the upper quartile of the within-subjects MSE ratio (after subtracting out 1). This improvement is due to a large increase in within-subject variability in beta-coefficients after 300 ms. Our results therefore show that ERP reliability is signal dependent because (1) both ERPs and $R^2(t)$ started to be reliable (high ICC) at the beginning of the N170 component, (2) ICC inter-quartile range of ERPs were very large at the beginning and the end of the N170 component and (3) MSE ratios can be improved by restricting the analysis to the early time window containing the N170.

### Sources of low reliability

Rousselet et al. (2007) measured the inter-trial variability of ERP to faces and other stimuli by measuring the range of temporal lags giving rise to a maximum cross-correlation between the mean ERP, clipped between 100 and 300 ms, and individual trials. This analysis of cross-correlation lag is similar to our measure, except that their analysis examined inter-trial, rather than inter-session, or inter-subject similarity. Using range of lag (max minus min) as a measure of inter-trial variability for each observer, and bootstrap statistics, they demonstrated that one of their observers was significantly less variable than all of the others. Such inter-trial variability between subjects is one possible explanation of some of our results: all subjects showed reliable ERP / $R^2(t)$ except for DVD for ERPs only and MLR for $R^2(t)$ only. The discrepancy ERP / $R^2(t)$ could thus be explained assuming that in DVD trials to faces/noise stimuli were highly variable while these variations were still influenced by the phase manipulation; by contrast for MRL, trials to faces / noise stimuli were stable while not very sensitive to the stimulus phase manipulation.

### Source of individual differences in shape

Individual differences could not have been time of day or time since last meal because we kept that constant (Geisler and Polich, 1992). Head size might have been a factor, although we do not have enough subjects to test that (Gregori et al., 2006). Correlations between retinal conductivity and VEP suggests that such correlations may also exist with ERP and EEG- regression-models measured during object-perception, especially given the difficulty of identifying noisy faces (Wright et al, 1985). Future study may reveal the role of both trivial (head-size) and non-trivial factors (brain anatomy). Another interesting source of differences in EEG shape would be variation in information processing. The best way to explore that hypothesis is to use EEG shape to predict behavioral performance during information processing tasks. A great deal has been done for EEG peaks like the P300, and for EEG spectra. And a great deal needs to be done for EEG during object based perception, especially for EEG shape-based metrics. However, as the next section points out, reliability studies like those proposed here can be used to select the most promising metrics.

### Prediction of individual differences in behavior

Few studies (Gauthier et al., 2003; Jacques and Rossion, 2007; Rossion et al., 1999; Schyns et al., 2007) to date have correlated behavioral performance with the N170, and yet these studies have produced very different results (Table 1). Schyns et al. (2007) obtained a very strong correlation between N170 latency and a behavioral measure based on reverse correlation. Gauthier et al. (2003) obtained a very weak correlation between a behavioral measure of expertise for car recognition, and the difference in amplitude between N170 for faces and for cars. Finally, Rossion et al. (1999) examined the difference in both amplitude and latency of the N170 between upright and inverted faces; they obtained very weak correlations between these 2 measures and behavioral measures. However, as Vul et al. (2009) had pointed out for fMRI data, low test–retest reliability of one measurement may introduce noise that can mask its true correlation with another measure. The estimated correlation between variables X and Y, $r_{X^\wedge Y^\wedge}$ is limited by the true correlation $r_{XY}$ and the reliability of each variable, $ICC_X$ and $ICC_Y$, respectively (Ghiselli, 1964):

$$r_{\overline{X}\overline{Y}} = r_{XY}\sqrt{ICC_X ICC_Y}. \qquad (4)$$

Given a true correlation between X and Y of 1 and an ICC of 1 for a behavioral variable Y, the maximum observable correlation is:

$$maxr_{\hat{X}\hat{Y}} = \sqrt{ICC_X}. \qquad (5)$$

In other words, the weaker correlation obtained by Gauthier et al. (2003) may have been caused by greater noise in the amplitude difference between N170 for faces and for cars. Similarly, the weaker correlations obtained by Rossion et al. (1999) may have been due to

**Table 1**
$N170_{faces}$ and $N170_{noise}$ refer to conventional peak-based components for face and noise stimuli respectively. $R^2(t)_{envelope}$ gives the maximum $R^2$ across electrodes, separately for every time point. A and L are peak amplitude and latency, respectively; and G is the latency giving 50% of the cumulative $R^2(t)$ or $R^2(t)_{envelope}$ integral, which ends at either 500 or 300 ms. Where available, $r_{X^\wedge Y^\wedge}$ is a reported correlation between the variable and some other variable. Gauthier et al. (2003) measured $N170_{cars\text{-}faces}$, not $N170_{faces\text{-}noise}$. Rossion et al. (1999) measured $N170_{faces\text{-}inverted\_faces}$.

| Variable | | | | | |
|---|---|---|---|---|---|
| Signal | Measure | ICC | max $r'_{XY}$ | $r'_{XY}$ | Study |
| $N170_{faces}$ | A µV | 0.72 | 0.85 | | |
| | L ms | 0.82 | 0.91 | 0.81 | Schyns et al. (2007) |
| $N170_{noise}$ | A µV | 0.66 | 0.81 | | |
| | L ms | 0.84 | 0.92 | | |
| $N170_{faces\text{-}noise}$ | A µV | 0.08 | 0.28 | 0.35 | Gauthier et al. (2003) |
| | | | | 0.32 to 0.52 | Rossion et al. (1999) |
| | L ms | −0.20 | 0 | 0.22 | Rossion et al. (1999) |
| $R^2(t)$ | G, 500 ms | 0.66 | 0.81 | | |
| | G, 300 ms | 0.89 | 0.95 | | |
| $R^2(t)_{envelope}$ | G, 500 ms | 0.86 | 0.93 | | |
| | G, 300 ms | 0.96 | 0.98 | | |

the greater noise in the amplitude and latency differences between the N170 for upright and inverted faces. We can test this by assuming that the true correlation in either study was 1, and that ICC for the behavioral measures were also both 1. Using the above equation, we can then measure the best possible correlation measurable in Schyns et al. (2007). We can also use the ICC of $N170_{faces-noise}$ as a rough estimate for the ICC of $N170_{faces-cars}$. The results are shown in Table 1. Despite differences in our experiment and theirs, one can see that our predictions based only on reliability are good. Also included in the table are a few other EEG metrics, based on both ERP peaks, and on aspects of the regression model. $R^2(t)_{envelope}$ refers to the maximum $R^2$ across channels, calculated separately for each time-point. Integration for $R^2(t)$ and $R^2(t)_{envelope}$ (measure 'G' in Table 1) refers to time required to cumulate a fixed proportion of the total integrated $R^2$ . Given our results on the time course of ICC, we decided to vary the time-window that determines total integration. We expected integration time to be more reliable for a time window of 300 ms compared to 500 ms. This prediction held up. As one can see, a wide variety of EEG metrics would be very useful for predicting individual differences in behavior. As far as we know, only $N170_{faces-noise}$ latency and amplitude, and $N170_{faces}$ latency, have been used for this purpose.

## Conclusion

For the first time, we present results of test–retest reliability of the 'shape' of the ERP time-course and the first test-retest reliability results for single-trial ERP analysis. Our results show that both ERPs and single trial analyses are highly reliable with shapes that correlate very well (r ~ 0.9) with almost no change in the temporal dynamic (lag ~4/6 ms). In addition, we observed that relative changes in amplitude were also very reliable around regions / periods of interest (e.g. ICC around the N170 component ~ 0.9). Altogether, these results demonstrate that (1) individual brain electrical evoked responses acts as neural signatures of particular events and (2) ERP grand averages do not reflect individuals' brain dynamics.

Supplementary materials related to this article can be found online at doi:10.1016/j.neuroimage.2011.06.052.

## References

Bland, J., Altman, D., 1996. Statistics notes: measurement error and correlation coefficients. BMJ 313, 41–42.

Burgess, A., Colborne, B., 1988. Visual signal detection. IV. Observer inconsistency. Journal of the Optical Society of America. A 5 (4), 617–627.

Daugman, J., 2003. The importance of being random: statistical principles of iris recognition. Pattern Recognition 36 (2), 279–291.

Delorme, A., Makeig, S., 2004. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. Journal of Neuroscience Methods 134 (1), 9–21.

Delorme, A., Sejnowski, T., Makeig, S., 2007. Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. NeuroImage 34 (4), 1443–1449.

Gaspar, C., Rousselet, G., 2009. How do amplitude spectra influence rapid animal detection? Vision Research 49 (24), 3001–3012.

Gaspar, C., Bennett, P., Sekuler, A., 2008. The effects of face inversion and contrast-reversal on effciency and internal noise. Vision Research 48 (8), 1084–1095.

Gasser, T., Bacher, P., Steinberg, H., 1985. Test–retest reliability of spectral parameters of the EEG* 1. Electroencephalography and Clinical Neurophysiology 60 (4), 312–319.

Gauthier, I., Curran, T., Curby, K., Collins, D., 2003. Perceptual interference supports a non-modular account of face processing. Nature Neuroscience 6 (4), 428–432.

Geisler, M., Polich, J., 1992. P300 and individual differences: morning/evening activity preference, food, and time-of-day. Psychophysiology 29 (1), 86–94.

Ghiselli, E., 1964. Theory of Psychological Measurement. McGraw-Hill.

Gold, J., Bennett, P., Sekuler, A., 1999. Identification of band-pass filtered letters and faces by human and ideal observers. Vision Research 39 (21), 3537–3560.

Gregori, B., Pro, S., Bombelli, F., Riccia, M., Accornero, N., 2006. Vep latency: sex and head size. Clinical Neurophysiology 117 (5), 1154–1157.

Husk, J., Bennett, P., Sekuler, A., 2007. Inverting houses and textures: investigating the characteristics of learned inversion effects. Vision Research 47 (27), 3350–3359.

Jacques, C., Rossion, B., 2007. Early electrophysiological responses to multiple face orientations correlate with individual discrimination performance in humans. NeuroImage 36, 863–876.

Jain, A., 2007. Biometric recognition. Nature 449 (9), 38–40.

Kiebel, S.J., Friston, K.J., 2004. Statistical parametric mapping for event-related potentials I: generic considerations. NeuroImage 22, 492–502.

Kuefner, D., de Heering, A., Jacques, C., Palmero-Soler, E., Rossion, B., 2010. Early visually evoked electrophysiological responses over the human brain (p1, n170) show stable patterns of face-sensitivity from 4 years to adulthood. Frontiers in Human Neuroscience 3.

Laming, D.R.J., 1968. Information Theory of Choice Reaction Time. Wiley, NewYork.

Lauritzen, L., Jørgensen, M., Michaelsen, K., 2004. Test–retest reliability of swept visual evoked potential measurements of infant visual acuity and contrast sensitivity. Pediatric Research 55 (4), 701.

Makeig, S., Debener, S., Onton, J., Delorme, A., 2004. Mining event-related brain dynamics. Trends in Cognitive Sciences 8 (5), 204–210.

Morgan, H., Klein, C., Boehm, S., Shapiro, K., Linden, D., 2008. Working memory load for faces modulates P300, N170, and N250r. Journal of Cognitive Neuroscience 20 (6), 989–1002.

Oppenheim, A., Lim, J., 1981. The importance of phase in signals. Proceedings of the IEEE 69 (5), 529–541.

Pernet, C.R., Chauveau, N., Gaspar, C., Rousselet, G.A., 2011. LIMO EEG: a toolbox for hierarchical LInear MOdeling of EletroEncephaloGraphic data. Computatioanl Intelligence and Neuroscience. ID 831409. .

Philiastides, M.G., Sajda, P., 2006. Temporal characterization of the neural correlates of perceptual decision making in the human brain. Cerebral Cortex 16 (4), 509–518.

Philiastides, M.G., Ratcliff, R., Sajda, P., 2006. Neural representation of task diffculty and decision making during perceptual categorization: a timing diagram. The Journal of Neuroscience 26 (35), 8965–8975.

Piotrowski, L.N., Campbell, F.W., 1982. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. Perception 11 (3), 337–346.

Ratcliff, R., McKoon, G., 2008. The diffusion decision model: theory and data for two-choice decision tasks. Neural Computation 20, 873–922.

Rossion, B., Delvenne, J., Debatisse, D., Goffaux, V., Bruyer, R., Crommelinck, M., Gu´erit, J., 1999. Spatio-temporal localization of the face inversion effect: an event-related potentials study. Biological Psychology 50 (3), 173–189.

Rousselet, G., Pernet, C., 2011. Quantifying the time course of visual object processing using ERPs: it's time to up the game. Frontiers in Psychology 2, 107. doi:10.3389/fpsyg.2011.00107.

Rousselet, G., Mace, M., Fabre-Thorpe, M., 2004. Animal and human faces in natural scenes: how specific to human faces is the n170 erp component? Journal of Vision 4 (1), 13–21.

Rousselet, G., Husk, J., Bennett, P., Sekuler, A., 2007. Single-trial EEG dynamics of object and face visual processing. NeuroImage 36 (3), 843–862.

Rousselet, G., Pernet, C., Bennett, P., Sekuler, A., 2008a. Parametric study of eeg sensitivity to phase noise during face processing. BMC Neuroscience 9, 98.

Rousselet, G., Husk, J., Bennett, P., Sekuler, A., 2008b. Time course and robustness of erp object and face differences. Journal of Vision 8 (3), 1–18 URL http://journalofvision.org/8/12/3/ (12).

Rousselet, G., Husk, J., Pernet, C., Gaspar, C., Bennett, P., Sekuler, A., 2009. Age-related delay in information accrual for faces: evidence from a parametric, single-trial eeg approach. BMC Neuroscience 10 (1), 114 URL http://www.biomedcentral.com/1471-2202/10/114.

Rousselet, G., Gaspar, C., Pernet, C., Husk, J., Bennett, P., Sekuler, A., 2010. Healthy aging delays scalp eeg sensitivity to noise in a face discrimination task. Frontiers in Psychology 1, 12.

Salinsky, M., Oken, B., Morehead, L., 1991. Test-retest reliability in EEG frequency analysis. Electroencephalography and Clinical Neurophysiology 79 (5), 382–392.

Schyns, P., Petro, L., Smith, M., 2007. Dynamics of visual information integration in the brain for categorizing facial expressions. Current Biology 17 (18), 1580–1585.

Segalowitz, S., Barnes, K., 1993. The reliability of ERP components in the auditory oddball paradigm. Psychophysiology 30 (5), 451–459.

Segalowitz, S., Santesso, D., Murphy, T., Homan, D., Chantziantoniou, D., Khan, S., 2010. Retest reliability of medial frontal negativities during performance monitoring. Psychophysiology 47 (2), 260–270.

Shrout, P., Fleiss, J., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86 (2), 420–428.

Taylor, M., Batty, M., Itier, R., 2004. The faces of development: a review of early face processing over childhood. Journal of Cognitive Neuroscience 16 (8), 1426–1442.

Tello, C., De Moraes, C., Prata, T., Derr, P., Patel, J., Siegfried, J., Liebmann, J., Ritch, R., 2010. Repeatability of short-duration transient visual evoked potentials in normal subjects. Documenta Ophthalmologica 120 (3), 219–228.

Van Rijsbergen, N.J., Schyns, P.G., 2009. Dynamics of trimming the content of face representations for categorization in the brain. PLoS Computational Biology 5 (11) Nov.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on Psychological Science 4 (3), 274–290.

Wright, C., Williams, D., Drasdo, N., Harding, G., 1985. The influence of age on the electroretinogram and visual evoked potential. Documenta Ophthalmologica 59 (4), 365–384.