Breakthrough Percepts of Famous Faces

Abdulmajeed Alsufyani[1] , Omid Hajilou[2], Alexia Zoumpoulaki[3], Marco Filetti[4], Christopher J. Solomon[5], Stuart J. Gibson[5], Roobaea Alroobaea[2], Howard Bowman[2,6]

1 Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, KSA.

2 School of Computing, University of Kent, Canterbury, Kent, UK.

3 School of Psychology, Cardiff University, Park Place, Cardiff, UK

4 Helsinki Institute for Information Technology, University of Helsinki, Helsinki, Finland

5 School of Physical Sciences, University of Kent, Canterbury, Kent, UK.

6 Department of Psychology, University of Birmingham, Birmingham, UK

Abstract

Recently, we showed that presenting salient names (i.e. a participant's first name) on the fringe of awareness (in Rapid Serial Visual Presentation) breaks through into awareness, resulting in the generation of a P3, which (if concealed information is presented) could be used to differentiate between deceivers and non-deceivers (Bowman et al., 2013; Bowman, Filetti, Alsufyani, Janssen, & Su, 2014). The aim of the present study was to explore whether face stimuli can be used in an ERP-based RSVP paradigm to infer recognition of broadly familiar faces. To do this, we explored whether famous faces differentially break into awareness when presented in RSVP and, importantly, whether ERPs can be used to detect these 'breakthrough' events on an individual basis. Our findings provide evidence that famous faces are differentially perceived and processed by participants' brains as compared to novel (or unfamiliar) faces. EEG data revealed large differences in brain responses between these conditions.

Key words: Familiarity, famous faces, EEG/ERP, P3, RSVP, Time-frequency analyses, Deception Detection.

The Fringe-P3 Method: There has been a sustained body of research focussed on developing EEG-based Concealed Information Tests, e.g. Labkovsky and Rosenfeld (2012). The objective in this work has been to provide a means by which a suspect's familiarity with a piece of incrminiating infromation can be demonstrated using EEG. The methods proposed to do this typically present a series of stimuli and monitor the EEG for a distinct evoked response pattern (usually a larger P3), which by the logic of the procedure would indicate familiartiy with a probe stimulus marking the incriminating information.

Bowman et al. (2013) and Bowman, Filetti, Alsufyani, Janssen, and Su (2014) introduced the Fringe-P3 method as a concealed information test, and showed that presenting stimuli on the fringe of awareness (in Rapid Serial Visual Presentation) provides countermeasure-resistant detection of familiarity with an individual's own identity. The significance of this technique is that, during Rapid Serial Visual Presentation (RSVP), salient stimuli (e.g. a participant's first name) break through into awareness, generating a P3 component, which (if concealed information is presented) could be used to differentiate between deceivers and non-deceivers. Non-salient stimuli, on the other hand, remain unreported (i.e. (sub)limina[1]), making it difficult for a participant to use a strategy (e.g. apply countermeasures) to confound the method.

The Fringe-P3 method, then, offers the possibility of a reliable Event Related Potential (ERP) concealed Information test that is resistant to the countermeasures that have confounded previous methods Rosenfeld et al (2004). However, to date, the method has only been demonstrated on a class of stimuli that could be expected to give big familiarity responses, own-name (Bowman et al, 2013 & 2014). To demonstrate the

[1] The exact nature of the experience of unreported items remains debated (Bowman et al, 2013). Accordingly, we bracket the "sub" to reflect the possibility that unreported items may, in some sense, be experienced, but not sufficiently to be recalled; see Block's notion of phenomenological awareness (Block, 2007) .

broader applicability of the Fringe-P3 method, there is a need to demonstrate that the "break-through" into awareness phenomenon that underlies it can be elicited and detected with a range of stimulus types. Making a first step in this direction is the main aim of this paper.

Specifically, the paper considers whether face images can be used in an ERP-based RSVP paradigm to infer recognition of familiar faces. We do this by assessing whether famous faces differentially break into awareness when presented in RSVP and, importantly, whether we can detect these "breakthrough" events on an individual basis using EEG (note, a concealed information test would have to be applied to individuals; thus inference at that level, and not just the group level, needs to be assessed). To answer these questions, an ERP experiment was conducted, following a protocol similar to that used in Bowman et al. (2013), with the exception that there were no instructions to conceal (or to lie about) the Probe (which for us is a famous face). In other words, there was no explicit task associated with the famous faces, and participants were not informed of their presence. If we can show that the Fringe-P3 method can work with famous faces, it will make an important step towards the use of the method to show familiarity with crime or terrorist compatriots. The work presented here also provides a second, more theoretical, contribution. This is the identification of a striking theta burst for specifically famous faces that is phase locked to the eliciting stimulus, and is remarkably consistent across participants.

A precedent for our work is the important research of Ganis and Patnaik (2009), who presented famous faces in RSVP. Their work, though, was purely behavioral. The research presented here uses EEG to provide an EEG marker of the brain detecting a famous face in RSVP; no such RSVP marker has been previously reported.

Face-related Components: Our ERP findings will be informed by a number of components that have previously been identified in face-related EEG studies. We highlight these previous findings here. Indeed, ERPs have been used extensively to investigate neural activity associated with face perception. Several studies have reported that face perception and recognition is associated with an enhanced negativity called the N400f (also referred to as the face-N400) (Bentin & Deouell, 2000; Eimer, 2000; Touryan, Gibson, Horne, & Weber, 2011a) and a late positivity that is referred to as the P300 (Bentin & Deouell, 2000; Meijer, Smulders, Merckelbach, & Wolf, 2007) or P600 (Eimer, 2000; Henson et al., 2003). The N400f is a negative deflection elicited in the ERP within ~250 ms to 500 ms post-stimulus, whereas the P600 is a positive deflection that can be observed between 500 ms and 800 ms post-stimulus.

Bentin and Deouell (2000) and Eimer (2000) investigated the cognitive process underlying recognition of familiar and unfamiliar faces. They found that ERPs elicited by familiar faces were more negative than those elicited by unfamiliar faces within the time window between ~250 ms and ~500 ms post-stimulus of the N400f. In addition to this increased negative deflection, a larger P600 late positivity (~500 ms to 800 ms post-stimulus) for familiar compared to unfamiliar faces, has been observed Eimer (2000), Henson et al. (2003), and Touryan, Gibson, Horne, and Weber (2011b).

Faces and Concealed Information Tests: In the context of forensic science (e.g. the Concealed Information Test (CIT) or Guilty Knowledge Test), the fact that familiar faces elicit distinct brain responses suggests the possibility of employing ERPs to detect whether a suspect is familiar with a particular person's face (e.g. the victim, or members of a terrorist ring). Related studies by Lefebvre, Marchand, Smith, and Connolly (2007) and Meijer et al. (2007) demonstrated that faces can be used effectively as stimuli in an ERP-based Concealed Information Test (CIT). However,

both of these studies were based on the classic P3 oddball paradigm, which has been found to be vulnerable to countermeasures (Rosenfeld, Soskins, Bosh, & Ryan, 2004). As previously discussed, there is evidence that the Fringe-P3 concealed information test is robust against countermeasures. The results reported in this paper make a key step towards a Fringe-P3 based concealed information test for faces.

The following sections describe the experiments and the EEG data analysis methods; then we present the results; and finally, we discuss our findings.

Experiment

Participants

Fourteen participants (age range 19-24, 6 male, 8 female) participated in the experiment. All participants were students at the University of Kent, right-handed, free from neurological disorders and had normal or corrected-to-normal vision. Only native English speakers participated in the experiment. All participants signed a consent form before participating in the experiment. The study was advertised publicly and each participant was paid 10 pounds (GBP) for participating. The Sciences Research Ethics Advisory Group at the University of Kent approved the study.

Stimuli and Experimental Design

Stimuli were presented using the Rapid Serial Visual Presentation (RSVP) technique. RSVP streams were presented on a 20-inch LCD screen with a refresh rate of 60 Hz and a resolution of 1600x1200. The screen was placed at a distance of 60~80 cm from the participant. Stimuli were presented using the Psychophysics toolbox version 3 running under Matlab 2012. The Stimulus Onset Asynchrony (SOA) was 133 ms. We presented all stream items at the same screen location.

Each RSVP trial consisted of 18 280*320 pixel grey photographs of famous and unfamiliar faces. All photographs showed frontal views of the faces; see Figure A.1.

Each RSVP trial contained one critical stimulus and 17 random (unfamiliar) faces used as (filler) distractors. There were three categories of critical stimuli: (1) Probes: faces of famous people from different fields such as, politics, entertainment and sport (5 faces of famous people were presented: Nelson Mandela, Barack Obama, Margaret Thatcher, David Beckham and Angelina Jolie); (2) Irrelevants: random faces that were unknown to the participants (5 faces of unfamiliar people, selected from the database of distractor faces); (3) Target: an irrelevant face, which participants were instructed to respond to (i.e. that are task-relevant). The Target-task was included to ensure participants maintained attention on the RSVP items.

Each of the five critical stimuli (faces) in the Probe and Irrelevant categories was repeated 15 times, resulting in 75 Probe trials and 75 Irrelevant trials. The (same) Target face was presented on 75 trials. Thus, there were 225 RSVP trials in total. The experiment was divided into 5 blocks. Each block comprised 45 trials; 15 Probe trials, 15 Irrelevant trials, and 15 Target trials. In each block, only one Critical item from each category (Probe, Irrelevant and the Target face) was presented; each of which was displayed 15 times. The order in which the critical stimuli were presented within blocks was randomized.

The position of the critical face within each stream was selected pseudo randomly, so that it had equal probability of appearing in the 5th position (earliest) through to the 9th position (latest). These positions were chosen to prevent beginning or end of stream effects overlapping with the ERP response to the critical item. For each RSVP trial, there were also starting and finishing items. The latter was either --------- or =========, selected at random, and remaining on screen for 133 ms. It was followed by a question asking the participant which of the two alternatives was shown. This end of stream item was included in order to maintain attention on all the stream's stimuli.

The starting item was XXXXXXXXX, which was presented for 800 ms, to position the participant's focus on the stimulus presentation area. The starting and finishing items were presented as photographs. Part of an RSVP stream of face stimuli is shown in Figure 1.

The photographs of unfamiliar faces (distractors) were obtained from an online database of neutral faces (524 faces) developed by Minear and Park (2004). The Irrelevant and Target faces were chosen randomly from the database of distractors. Photographs of famous faces were collected from different online websites. They were processed manually using graphics software to obtain similar physical parameters (i.e. brightness and contrast) to those of unfamiliar faces. All photographs were centred by aligning the eye-line to the same horizontal position. The background of each photograph was selected (i.e. highlighted using the free-select tool), and replaced with a standard grey background colour (#e7e7e7 Hex colour). Each Photograph was converted to grey-scale, and a 'blur' tool was used to smear the edges of the head/hair, to avoid any sharp transitions between the face and its background. The contrast of each photograph was reduced, wherever necessary. We excluded from the experiment any faces with significant facial expressions and faces of people wearing glasses or hats.

Moreover, to ensure that there were no significant differences between the visual properties (i.e. brightness and contrast) of famous and unfamiliar faces, we statistically analyzed the pixel intensities of each photograph in the Probe (famous faces) and Irrelevant (unfamiliar faces) categories. We calculated the first four moments (mean, variance, skewness, kurtosis) of pixel intensities of each photograph in both groups. Note that, in a gray scale image, each pixel is characterized by an intensity value (256 different possible intensities), representing the brightness of that pixel. We

ran a two sample independent t-test to compare pixel intensity properties between famous and unfamiliar faces. At an alpha level of 0.05, no significant differences were found between physical properties of faces in the two categories. Results of these statistical analyses are shown in Table 1.

Tasks

Prior to the start of the experiment, participants were given a Target face, and instructed to respond 'Yes' or 'No' at the end of each trial, when they were asked the question, 'Did you see the Target face?'. This question followed the finishing item question (described above). Participants had to use a numeric keypad with their right hand, pressing '1' for Yes and '2' for No. The Target face was a random face, and was not (pre-experimentally) familiar to the participants. Participants were not told about the presence of the Probes (the famous faces). They were also informed that after each block, there would be a recognition test, explained in the next section. Participants were also instructed to keep their eyes fixated on the center of the screen, and to avoid eye movements during each trial.

Before the EEG recording, we ran a training session of 20 trials with a recognition test. The aim of this session was only to familiarize participants with the presentation of stimuli, and also to make sure that they were able to identify the Target. During this session, we did not present the Probe stimuli (the famous faces) or any faces of other famous people. Only Target trials (trials containing the Target face) were included.

Recognition Tests

This test was conducted to explore participants' memory, and to make sure that the Target and the Probe had indeed been recognized. There were five categories of faces used for the recognition test: Target, Probe, Irrelevant, Un-presented famous face

(faces of famous people that were not presented at all in RSVP streams during the experiment) and distractors. Apart from the Target (since there was only one Target face across the entire experiment), each category was composed of five different faces, as the experiment consisted of 5 blocks. The faces of the distractor category were chosen randomly from our distractors database, and may or may not have appeared in the experiment. Each distractor face had the same probability (0.032) of being presented in any given trial (17 distractor faces were presented in each trial) as any of the possible 524 (number of faces in the database) distractor faces.

At the end of each block, participants were presented with five faces (one from each category), and asked to give a confidence rating of how often each one of them had appeared in the preceding block. The ratings were on a scale of 1 to 5, where 1 meant 'Not appeared at all' and 5 meant 'Appeared very often'.

In this test, our hypothesis was that when presenting items in rapid succession (as is the case in our RSVP stream here), only items that capture attention because of salience (i.e. familiarity) would be consciously reported at the end of a block.

Data acquisition

We recorded data using a BioSemi ActiveTwo system (BioSemi, Amsterdam, The Netherlands). Data were filtered at recording, with a low-pass of 100Hz. Electroencephalographic (EEG) data were recorded at the Fz, Cz, P3, Pz, P4, Oz, A1 and A2 electrodes based on the standard 10-20 system. During recording, data were referenced to a ground formed from a common mode sense (CMS) active electrode and driven right leg (DRL) passive electrode (see http://www.biosemi.com/faq/cms&drl.htm). The Electrooculograms (EOG) generated from blinks and eye movements were recorded from the participant's left and right

eyes, using two bipolar HEOG and VEOG electrodes. We kept impedances below 10 kOhms. During acquisition, data were digitized at 2048Hz.

Analysis of EEG data

Recorded data were analysed with EEGLAB version 12.0.2.4b under Matlab 2013a (Delorme & Makeig, 2004). At analysis, we resampled the data at 512Hz. We filtered the data with a low-pass of 45Hz and high-pass of 0.5 Hz. In order to remove the Steady State Visually Evoked Potentials (SSVEP) oscillation set-up by the stream presentation, we filtered out 7 - 9 Hz band. This is the same filtering approach we took in Bowman et al (2014), which we further justify in appendix B. We re-referenced data offline to the average of the combined mastoids (electrodes A1 and A2). ERPs were generated by separately averaging all trials in each condition: the Target, the Probe, and the Irrelevant. EEG data were epoched using 100 ms before the onset of a critical item as a baseline. EEG trials[2] were 1.2sec long.

Eye blinks were detected by marking any activity below -100 μV or above +100 μV in the EOG channels. For the scalp channels, we automatically rejected any trials containing electrical activity below -50 μV or above +50 μV, in a time window from -100 ms to 1200 ms, with respect to the critical stimulus onset. We also performed a manual inspection to verify that the rejected trials were accurately detected.

The number of trials remaining after artefact rejection, per condition, ranged between 59 and 73; Target (M = 68.91, SD = 5.64); Probe (M = 71.72, SD = 3.53); Irrelevant (M = 71.57, SD = 4.21); this is out of 75 trials recorded for each condition. None of the participants were excluded from the analysis due to excessive artefacts (e.g. eye blinks).

[2] We use the term trial rather than epoch to maintain consistency with ours' and others' previous work.

Data Analyses

Our analyses in the present study were performed at the ERP level (as is typical in ERP-based deception detectors) as well as in the time-frequency domain. The goal of the analyses was to compare between EEG responses to famous and unfamiliar faces. This was done on a participant-by-participant basis (on participant-level ERPs) as well as at the group level.

ERP level analyses. We used mean amplitude measurements, which are more robust against high frequency noise (Luck, 2005, p. 234). The mean amplitude measurement has been used in previous ERP studies comparing famous and unfamiliar faces (Curran & Hancock, 2007; Eimer, 2000; Picton et al., 2000; Touryan et al., 2011b). As noted in the introduction, there are two ERP patterns that have been found to be sensitive to the recognition of familiar faces: an enhanced negativity, called the N400f (~250-500 ms), and a late positivity (~500-800 ms) (Bentin & Deouell, 2000; Curran & Hancock, 2007; Eimer, 2000; Touryan et al., 2011b). In agreement with Eimer (2000), Trenner, Schweinberger, Jentzsch, and Sommer (2004), we refer to the late positivity as a P600.

As Brooks, Zoumpoulaki, and Bowman (2017), Kilner (2013), and Kriegeskorte, Simmons, Bellgowan, and Baker (2009a) have pointed out, selecting time windows based on where an effect being tested is largest generates a reporting bias, inflating false positive rates (i.e. it increases the probability of detecting an effect when none exists or, in other words, increases the Type I error rate). To avoid this problem, our time windows (N400f and P600) of interest were selected based on an orthogonal contrast; see Brooks et al. (2017) and Friston, Rotshtein, Geng, Sterzer, and Henson (2006) for a justification of this approach. In other words, the window placement needs to be made independently of the contrast that is tested (Brooks et al., 2017; Kilner,

2013; Kriegeskorte, Simmons, Bellgowan, & Baker, 2009b). Therefore, to identify the parameters of the time window of interest for the group level analyses, we selected the time window using an aggregated grand average of trials (AGAT) generated from averaging all the Probe and Irrelevant single trials across all participants. On the other hand, for the individual level analyses, a participant's ERP generated from all trials in Probe and Irrelevant conditions; which we call the aggregated ERP of trials (Brooks, Zoumpoulaki, & Bowman, 2017) was used. This aggregated average of trials approach is extensively investigated and justified in Brooks et al. (2017), in particular, it resolves problems of non-orthogonality arising from trial count asymmetry identified in (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). To defend ourselves most fully against criticisms of fishing for effects in the published literature, while we use the N400f and P600 terminology and concepts, window selection is fully data driven, but the AGAT (orthogonal contrast) approach to prevent inflating the type I error rate.

The size (mean amplitude) of the N400f, respectively P600, for each condition (Probe and Irrelevant) was quantified in two steps: (1) finding the N400f, respectively P600, time window from the aggregated grand average for the group level analyses or a participant's aggregated ERP of Probe and Irrelevant trials for the individual level analysis; (2) calculating the mean amplitude value for each ERP condition in the time window identified in step 1.

In more detail, for the group level analyses, the time window of interest was selected on the Aggregated Grand Average of Trials (AGAT). The size (mean amplitude) of the N400f, respectively P600, for each condition (Probe and Irrelevant) was quantified in two steps:

(1) We found the N400f, respectively P600, time window from the AGAT of all Probe and Irrelevant trials across all participants. For the N400f, the algorithm searched from

50 ms to 1200ms (which is the so called outer window) post-stimulus, to find the minimal 100ms interval average (which is the so called inner window). For the P600, the algorithm searched from 50 ms to 1200 ms (the outer window) post-stimulus, to find the maximal 100ms interval average. (Note, in both these cases, the outer window is the entire analysis segment, making it completely unconstrained).

The start and end of this minimal/maximal 100ms defined the N400f, respectively P600, and time window of interest for each condition.

(2) We calculated the mean amplitude value for each ERP condition in the time window identified in step 1.

However, for the individual level analysis, the aggregated ERP of all trials of Probe and Irrelevant conditions was determined for each participant. This aggregated ERP was then used to identify the time window of each component (N400f, respectively P600). For the N400f, the algorithm searched automatically from the lower boundary of the inner window of the N400f that was defined from the AGAT in the group level analyses to 1200ms post-stimulus, to find the minimal 100 ms interval average. On the other hand, for the P600, the algorithm searched from the lower boundary of the inner window of the P600 that was defined from the AGAT in the group level analyses to 1200ms post-stimulus, to find the maximal 100 ms interval average[3]. The start and end of this minimal/maximal 100 ms defined the N400f, respectively P600, time window of interest for each condition. After defining the time windows for each component, the mean amplitude within the given window was calculated separately for each condition (Probe and Irrelevant).

[3] For completeness, we also ran a P600 analysis with an outer search window from 50 ms to 1200 ms. This gives an even more unconstrained analysis, but for some participants picks up a positive deflection before the N400f, meaning that it is not in all cases truly a P600 analysis. Nonetheless, we can report that the fully unconstrained analysis (outer window from 50 ms to 1200 ms), only incurs a small reduction in statistical power relative to the findings we report shortly for the P600.

Statistical analyses of ERP data. A randomization (i.e. Monte Carlo permutation) test was used to determine whether the N400f/P600 evoked by the Probe was significantly different from that evoked by the Irrelevant. Specifically, a randomization was applied in order to generate participants' null hypothesis distributions; see Figure 4 for such distributions. Before applying the test, the difference between the mean amplitude values of the N400f/P600 in the Probe and Irrelevant conditions was calculated. This resulted in two mean amplitude difference measures that were considered to be true observed values.

True_Observed _N400f = N400f Probe – N400f Irrelevant

True_Observed_ P600 = P600 Probe – P600 Irrelevant

The statistical significance of these two true observed values was assessed by computing randomised null hypothesis distributions. For each participant, two distributions were generated: one for the N400f and the other for the P600. Specifically, for each component (N400f/P600), each distribution comprised 1,000 randomized mean amplitude differences, using the randomization procedure described in (Bowman et al., 2013), and re-iterated here. That is, on each resampling, condition labels (for Probe and Irrelevant) are randomized, generating surrogate-Probe and surrogate-Irrelevant trial sets. Surrogate-Probe and surrogate-Irrelevant ERPs are then generated from these trial sets. From each of these, a mean amplitude is calculated. The difference of these (surrogate) mean amplitudes provides one (randomized mean amplitude difference) entry for the null hypothesis distribution (see Figure 4 for two example N400f null hypothesis distributions).

It is important to note that these values were calculated from the same random sample (i.e. in each resampling, a randomized mean amplitude difference was measured for both the N400f and the P600 component)—that is, these values were not calculated

in separate randomizations. A p-value was then determined for each component (by, in the usual way, calculating the proportion of null hypothesis data points above the true observed value), and so, each participant was assigned two p-values.

However, in real-life applications (e.g. lie detection), it would be important to generate a single measurement that could be used to judge whether a suspect was guilty or not. For that reason, as described in (Bowman et al, 2013), we used the Fisher combining procedure to calculate a joint p-value (across N400f and P300) for each participant (Hayasaka & Nichols, 2004). The Fisher procedure can be viewed as a non-parametric statistical inference method for handling a multivariate dependent measure. Parametric approaches would use Hotelling's T (bivariate case) or MANOVA. These standard parametric methods are not suitable for individual-level ERP analysis, because it is very difficult to robustly measure a variable of interest from a single trial, since it contains so much noise. Consequently, a resampling procedure that generates surrogate ERPs is required. The natural way to do this is with a permutation test, and the Fisher is a natural way to extend such a test to the multivariate case. The following subsection describes the Fisher procedure.

Combined analyses for ERP data. First, for each participant, the two (true observed) p-values obtained from the two null hypothesis distributions (N400f and P600) described above were used to calculate a true observed Fisher value, using the Fisher combining procedure. The method, which dates back to Fisher, was championed by Hayasaka and Nichols (2004) as a means to combine two separate dimensions to provide a single overall assessment of statistical significance, and it is the standard procedure we use to combine dimensions (e.g. Bowman et al., 2013; Bowman et al., 2014). Let P^{N400f} be the p-value of the N400f, and P^{P600} be the p-value of the P600; the true observed Fisher value was determined as:

$$True\_Observed\_Fisher\_Value = -2\log(P^{\boldsymbol{N400f}} \times P^{\boldsymbol{P600}})$$

As extensively explored in Bowman et al. (2013), the critical point about the Fisher method's formula is that the p-values are multiplied. This means that there are two particular combinations of constituent p-values where Fisher combining is especially effective (see the Appendices in Bowman et al., 2013). A) if there is good evidence under both N400f and P600 individually for their true observed effect (i.e. small p-values on both), a lot of evidence will accrue when the two are combined (i.e. a very small p-value will result). B) if one constituent p-value is very very small and the other is big, the Fisher combined p-value can also be very very small.

Critically though, a further randomization procedure is performed at the level of Fisher values; see next two paragraphs. In particular, it is certainly not the case that the product of constituent dimension true-observed p-values is used as the combined p-value, an approach that would dramatically inflate false positive rates.

Second, each value (denoted q) in each of the null hypothesis distributions (i.e. for N400f and P600) generated from one of the 1000 randomizations was converted to a p-value. That is, a p-value in each distribution (N400f or P600) was calculated as the proportion of randomized values (1000 randomized mean amplitude values present in each distribution) that were greater than q. This yielded 2,000 p-values (1,000 for each component – N400f and P600), or strictly, 1000 pairs of p-values, since the same randomization of trials generate an N400f and a P600 value. Let P_i^{N400f} be the p-value for the N400f of randomization i, and P_i^{P600} be the p-value for the P600 of the same randomization i. Then, the Fisher value of randomization i was defined using the following formula:

$$W_i^F = -2\log(P_i^{\boldsymbol{N400f}} \times P_i^{\boldsymbol{P600}})$$

where i ranges from 1 to 1000, because 1000 randomizations were performed. Applying this formula to all of the 2000 (or 1000 pairs of) p-values will yield 1000 randomized Fisher values, one for each randomization, which provides a distribution of Fisher values: see Figure 5 for a typical example. Finally, the combined Fisher p-value was determined by calculating the number of Fisher values (each from a particular randomization) that were greater than the true observed Fisher value, divided by 1000, as illustrated in Figure 5.

Importantly, we verified the validity of this Fisher combining procedure in Bowman et al (2013), where, in particular, we performed a false positive test of the method, with data generated under the null hypothesis with the correlation structure of EEG data. As required, the identified false positive rate was the alpha-level. Additionally, p-values, which is the measure being combined, is a standardized measure that normalizes amplitude differences between dimensions.

Time-frequency analyses. To analyze EEG data in the time-frequency domain, two transforms were used. The first transform calculated the average changes in the frequency power spectrum across all individual trials that are time-locked to the same stimulus; in EEGLAB, this is called an Event Related Spectral Perturbation (ERSP) (Delorme & Makeig, 2004; Makeig et al., 2002). The second transform, known as Inter-Trial Coherence (ITC) in EEGLAB, measures phase consistency between trials, determining in particular the extent to which individual trials are phase-locked at each time point and frequency range (Makeig et al., 2002; Scott Makeig, Debener, Onton, & Delorme, 2004).

A previous study by Bentin and Deouell (2000) reported that famous faces evoked ongoing oscillations from about 100 ms to 500 ms post-stimulus onset over parietal and occipital sites (see Figure C.1 in the Appendix). Similarly, the grand

average ERPs (presented in Figure 2) in our study will show that the Probe (the famous face) elicited a multi cycle oscillation pattern, which classic ERP analysis methods (i.e. Peak-to-Peak or base-to-peak) would not fully reflect or measure. For this reason, time frequency analyses (ERSP and ITC) were used here to measure this oscillation, with the work of Bentin and Deouell (2000) providing an a-priori precedent for our analysis.

Window placement for ERSP and ITC. Time frequency spectra were measured using orthogonal-contrast time windows. This window was placed based on an orthogonal-contrast. Similarly to the ERP analysis, the time window of interest that we used to measure the ERSP and ITC was identified based on an aggregated power/coherence. As pointed out earlier, the window placement needs to be made orthogonal to the contrast that is statistically tested (Kilner, 2013; Kriegeskorte et al., 2009). Therefore, for the group level analyses, the critical time window (highest 100ms in outer window from 50ms to 1200ms) for measuring ERSP/ITC was placed based on the aggregated grand average of trials, calculated using the average of power/coherence of all single trials from both Probe and Irrelevant conditions (across all participants). While, for the analyses by individual, a participant's critical time window for measuring ERSP/ITC was placed based on the aggregated power/coherence, calculated using a participant's average of power/coherence across both Probe and Irrelevant conditions. For example, the aggregated power of a given participant was calculated by taking the average of the changes in power in the frequency spectrum across all their individual trials in the Probe and Irrelevant conditions. The aggregated coherence was determined in a corresponding way. This aggregated power/coherence was then used to identify a time window of each transform (ERSP/ITC). In particular, an algorithm searched from 50ms ms to 1200 ms post-stimulus to find the maximum 100 ms interval average (power/coherence). After defining this orthogonal-contrast time windows, the

ERSP/ITC were measured separately in each condition (Probe and Irrelevant), as described in the following paragraph.

The EEGLAB time-frequency function (newtimef) (Delorme & Makeig, 2004) was used to calculate the ERSP and ITC of each condition. As the key comparison was between individual trials of the Probe and Irrelevant conditions, the inputs to the newtimef function were two matrices of size (time-points × number of trials), one for the Probe trials and the other for the Irrelevant trials. The outputs of the function were also two matrices; the first of these comprised the differences in power between Probe and Irrelevant, and the second comprised the differences in coherence between Probe and Irrelevant. From each of these two matrices, a single difference measurement was obtained for each transform (power and coherence) by taking the sum (across the identified, orthogonal contrast, time window) of all the values that were,

(1) greater than zero (on the assumption that high values of ERSP and ITC indicate the existence of evoked and induced activity that the procedure aimed to detect) and

(2) in a frequency range of 0.5–7 Hz.

The highest frequency was chosen as 7 Hz in order to avoid inclusion of the Steady State Visually Evoked Potentials (SSVEP) oscillation set-up by the RSVP stream presentation rate; the lowest frequency was set at 0.5 Hz, as this was the high pass filtering applied in the early stages of pre-processing the raw EEG data. This summation process resulted in two difference measures, one for power and the other for coherence. These two difference measures became the true observed values that were used later to calculate a p-value for each transform. ERSP and ITC were calculated using a Fast Fourier transform with a baseline correction from -100 ms to 0 ms.

Statistical analysis of ERSP and ITC. As in the ERP analysis, a randomization (Monte Carlo permutation) procedure was used to generate two null hypothesis distributions, one for power and the other for coherence, in both cases calculated by the summation process across the window selected under the orthogonal contrast just described. Using the true observed value obtained by the same summation process, a p-value was calculated for each transform. The Fisher combining procedure was then used to calculate a single p-value for each individual; this procedure was similar to that used to combine the p-values of the N400f and P600 (described in the previous subsection), except that here power and coherence were combined rather than the different components (N400f and P600).

Results

ERP Data

Figure 2 depicts the grand average ERPs of all conditions across the midline electrodes (Fz, Cz and Pz). The Target elicited a large P3. This was expected, since participants were explicitly instructed to detect the Target. However, our main comparison was between the Probe (famous faces) and Irrelevant (unfamiliar faces). As can be seen from the grand average ERPs of each of the midline electrodes, the Probe elicited a continuous oscillation pattern within a time window from 100 ms to 650 ms[4].

From within this oscillatory pattern, a large negative deflection followed by a smaller positivity was observed in a time window from 300 ms to 620 ms, with respect to the onset of the stimulus. Such a pattern was absent in the Irrelevant ERPs. As mentioned in the introduction, this negativity is referred to as an N400f (Bentin &

[4] Indeed, if one focusses specifically on the pattern from 50 to 700 ms, one can see a very accurate likeness of a three-cycle tapered sinusoid, except that it is increasingly skewed positively as one progresses towards 700 ms. The last author has been minded to call this pattern either a "drunken wavelet" or an "inebriated Mexican hat", with the drunkenness due to the skewing.

Deouell, 2000; Eimer, 2000; Touryan et al., 2011a). In line with Eimer (2000) and Touryan et al. (2011a), we refer to the positivity that follows the N400f as a P600.

Statistical Analysis

Analysis of ERP data at the group-level. Our group level statistical test was a paired t-test of the mean amplitudes of Probe N400f/P600 and Irrelevant N400f/P600 across all participants.

For the group level analyses, the size (mean amplitude) of the N400f (respectively P600), for each condition (Probe and Irrelevant) was quantified by finding the N400f (respectively P600), time window from the aggregated grand average of trials, which is constructed from all Probe and Irrelevant trials across all participants.

Within the N400f time range, the statistical test showed a highly statistically significant difference between the Probe and the Irrelevant at Fz [t (13) = -4.7560, p = 0.0003753, d= -1.768], Cz [t (13) = -6.1377, p = 0.00003557, d= -2.38] and Pz [t (13) = -6.6823, p = 0.000015, d=-2.39]. With regard to the late positivity (P600), the outcomes also revealed a significant group difference between the Probe and the Irrelevant at Fz [t (13) = 3.2376, p = 0.0065, d=1.317], Cz [t (13) = 3.4301, p = 0.0045, d=1.48], and Pz [t (13) = 5.5597, p = 0.0000923, d=2.033].

Per-individual mean amplitude values of N400f and P600 in the Probe and Irrelevant conditions at Pz are provided in Table 2. Table 3 presents the summary of the group level analyses of Probe N400f/P600 and Irrelevant N400f/P600 across the mid-line electrodes.

Analysis of ERP data by individual. As previously discussed in subsection Statistical Analysis of ERP Data, in order to statistically assess individuals' data, we used a randomization (i.e. Monte Carlo permutation) test, with a difference of mean amplitudes measure. We performed these analyses for the N400f and P600 at the Pz

electrode. This is justified under the orthogonal contrast logic, since the most extreme (smallest for N400f, largest for P600) 100 ms mean amplitude under the AGAT was in both cases at Pz (N400f: Fz: -0.7214, Cz: -1.5480, Pz: -2.1667; P600: Fz: 1.6884, Cz: 1.4898, Pz: 1.7053). This is in agreement with Kaufmann, Schulz, Grünzinger, and Kübler (2011).

Our critical test was between the Probe and Irrelevant ERPs. In particular, there were two contrasts, one comparing the N400f mean amplitude of the Probe and Irrelevant and another comparing the P600 mean amplitude of the Probe and Irrelevant. Using randomization tests with difference of mean amplitude measures, we generated two null hypothesis distributions for each participant: one for the N400f contrast and the other for the P600 contrast. This resulted in two p-values (one for each true observed value) for each participant. A single combined p-value was then determined for each participant using the Fisher combining procedure (as explained in subsection "Combined Analysis for ERP Data"). This combined p-value indicates whether ERP responses elicited by Probe (famous faces) were significantly different from those elicited by Irrelevant (unfamiliar faces) for that individual. We used an alpha level of 0.05 for all our statistical tests.

N400f. 71 % (10/14) of the participants showed a significant difference between the mean amplitude of the Probe and Irrelevant within an independently identified (orthogonal) 100 ms window selected in an (outer window) time range of 50-1200 ms. Six participants have p-values below 0.01 and four had slightly higher p-values, but still below our critical alpha level (0.05). Per-individual p-values are presented in Table 4 (column 2). Figure 3 depicts the ERPs at Pz for the Probe and Irrelevant conditions across all participants. Apart from Participant 11, all elicited clear negative deflections within a time window from ~50ms to ~1200 ms. The Probe of Participant 11 has no

clear negativity (relative to a zero baseline). The Probe is, to some extent, similar (or slightly larger) to the Irrelevant within the 100 ms critical time window (around 400 ms after the stimulus) selected from the aggregated ERP ($p = 0.453$). The N400f null hypothesis distribution for Participant 11 is shown in Figure 4 (right). The true observed value (N400f Probe minus N400f Irrelevant) was very small and fell around the distribution's center, which resulted in a large p-value. In contrast, the Probe of Participant 2 (one of the five participants whose data yielded the most significant differences: p-values < 0.001) elicited an enhanced negative peak (within the N400f time region), which resulted in a high Probe – Irrelevant mean amplitude difference and a highly significant p-value: $p < .001$. The corresponding null hypothesis distribution for Participant 2 is presented in Figure 4 (left).

P600. At an alpha level (0.05), only seven (50%) participants showed significant P600 Probe effects. Per-individual p-values for the P600 are provided in Table 4 (column 3). At an alpha level of 0.1, which is the typical level of significance used in P3-based deception detection studies (Dietrich, Hu, & Rosenfeld, 2014; Labkovsky & Rosenfeld, 2012; Rosenfeld et al., 2004), 9 participants (64%) showed a significant difference between Probe and Irrelevant. Interestingly, some participants who elicited significant N400f effects showed no evidence of the presence of the P600. As can be observed from Figure 3, Participants 6 and 9, whose data revealed significant N400f ($p < 0.019$), showed a similar pattern for Probe and Irrelevant in the P600 time region. As can be noted from Tables 2 and 4, for these participants, the P600 mean amplitudes were almost equal for the Probe and Irrelevant, resulting in high p-values; 0.365 and 0.496 for participants 6 and 9, respectively.

Combined analysis (N400f and P600). Per-individual combined Fisher values are presented in Table 4 (last column). Using the Fisher combining procedure, we were

able to combine the p-values of the N400f and P600 of each participant into a single joint p-value. For all participants (100%), the p-values were smaller than the critical significance threshold (0.05). Eight participants have p-values less than 0.001, and six have higher p-values, but still below the significance level 0.05. As can be noted from Table 4, the Fisher combined measure worked extremely well with our data. In particular, any participants who show strong effects in only one component (either N400f or P600) obtained small p-values under Fisher's approach. For example, Participant 11 has high N400f p-value (0.453) and relatively small P600 p-value (0.017); however, the Fisher combining method yielded a significant overall p-value (0.0.028).

Similarly, Participants 6, 9, and 14 have high P600 p-values (0.365, 0.496, and 0.491 respectively) and relatively small N400f p-values (0.011, 0.019, and 0.001 respectively). In spite of this, Fisher's method resulted in significant combined p-values (0.014, 0.033, and <0.001 respectively) for these participants. The Fisher (N400f and P600 combined) null hypothesis distribution for Participant 9 is presented in Figure 5. As discussed in subsection "Combined Analysis for ERP Data" and returned to in the "Discussion", Fisher combining is known to work well when one p-value is very very small, while the other is large (Bowman et al., 2013; Hayasaka & Nichols, 2004).

Time-frequency domain. As mentioned in subsection "Time-frequency analysis" in the "Methods" section, Event Related Spectral Perturbation (ERSP) and Inter-Trial Coherence (ITC) transforms were calculated across single trials of Probe (famous faces) and Irrelevant (unfamiliar faces). ERSP estimates the average changes in the power spectrum across all individual trials that are time-locked to the same stimulus (generating a time-by-frequency plot of change in power). ITC measures the phase consistency (synchronization) between individual trials at each time point and

frequency (Makeig et al., 2002; Ming et al., 2010). Such time-frequency analyses enable us to assess the multi-cycle oscillatory pattern, which for some participants starts as early as 50 ms post Probe onset.

The ERSP/ITC of each condition was calculated relative to a baseline window of -200 ms to 0 ms pre-stimulus (Probe or Irrelevant) onset. In the ERSP/ITC time-frequency plots, power and coherence at each frequency and time-point is indicated by a colour value. An increase in power/coherence is indicated by redness, while blueness indicates a decrease. Green areas indicate no significant changes in power/coherence. As mentioned earlier, at the group level, the critical time window (i.e. orthogonal-contrast time window) for measuring ERSP/ITC was placed based on the aggregated grand average of trials of power/coherence. Figure 6 depicts the ERSP results of the aggregated grand average of trials (AGAT) from both conditions (obtained from 14 participants).

It is worth noting that one reason for applying notch filtering between 7 and 9 Hz to remove the SSVEP set up by the stream presentation, was to obtain a cleaner ITC pattern; see Appendix B. Furthermore, we have performed time-frequency statistical analyses without the 7-9 Hz notch filter, while still finding analysis windows with the AGAT method. These analyses confirm significant differences between the ERSP of the Probe and that of the Irrelevant ($p = 0.0068$). The difference of the ITC of the Probe and Irrelevant was also statistically significant ($p = 0.0109$).

ERSP. Figure 7, top row depicts the ERSP results across all trials (obtained from 14 participants) of the Probe and Irrelevant conditions. The grand Probe-related power changes (relative to the baselining window) plot is presented in the first column of the figure, while the same for Irrelevant is presented in the second column. The difference between Probe and Irrelevant ERSPs is shown in the third column. Visual

inspection of the ERSP plot (in Figure 7, top row) of the Probe and Irrelevant conditions shows that, for the Probe, power increased in a time window between ~250 ms and 520 ms (post-stimulus onset) and over frequency bands from 0.5 Hz to 10 Hz, whereas in the Irrelevant condition, there was no significant change in power within the same time-window and over the same frequencies. The ERSP difference plot (ERSP Probe – ERSP Irrelevant, see Figure 7, top row, right column) shows a high difference in power (indicated by red) in a time window from 250 ms to 520 ms (post-stimulus onset) and over 0.5 Hz to 10 Hz frequencies. It is clear from these plots that there were no significant power fluctuations in frequencies higher than 10 Hz in any of the conditions. To confirm this visual impression, statistical tests were used to compare power changes in the Probe and Irrelevant conditions at both group and individual level.

As mentioned in subsection "Time-frequency analysis" of the "Methods" section, in order to obtain a single ERSP measure for each condition, the sum was calculated of all positive changes in power over frequency bands from 0.5Hz to 7Hz. By so doing, two single ERSP measures were obtained for each participant, one for the Probe and the other for the Irrelevant. To compare these measures at the group level, a two-tailed paired t-test was used. When applied to the orthogonal-contrast time window, the outcome confirms highly significant differences between the ERSPs of the Probe and the Irrelevant ($p < 10^{-7}$)[5]. This strong group effect was supported by per-individual analysis.

For the individual-level analyses, the statistical analysis revealed a significant difference between ERSPs of Probe and Irrelevant in all participants (100%). Nine

[5] To clarify again, this result is obtained with notch filtering to remove the SSVEP set-up by the stream. Running the same analysis without notch filtering gives a p-value of 0.0068, i.e. the effect remains.

participants had p-values below or equal to 0.001 and five participants obtained slightly higher p-values, but these were still below 0.05. Per-individual ERSP p-values (calculated within the 100 ms orthogonal-contrast time window) are presented in Table 5.

ITC. Figure 7 (bottom row) depicts the ITC results across all trials (obtained from 14 participants) of the Probe and Irrelevant conditions. The grand Probe-related coherence plot is presented in the first column of the figure, while the grand Irrelevant-related coherence plot is presented in the second column. The difference between ITCs of the Probe and Irrelevant conditions is shown in a difference plot in the third column. Visual inspection of the ITC plots of Probe and Irrelevant shows that the Probe produced higher ITCs in a time window between ~150 ms and 520 ms (post-stimulus onset) and over frequency bands from 0.5 Hz to 10 Hz. In the Irrelevant condition, there was no evident change in the ITC (relative to the baseline window) within the same time-window and over the same frequencies. As can be seen from the Probe ITC plot, the highest ITC (indicated by dark red in the plot; ITC ~> 0.4) was obtained over the N400f time period (~320–500 ms) and at frequencies from approximately 0.5 Hz to 8 Hz. The ITC difference plot (ITC Probe – ITC Irrelevant) shows high differences in the ITC within a time window from 150 ms–520 ms (post-stimulus onset) and over 0.5 Hz to 10 Hz frequencies.

As with the ERSP analysis, ITC changes in the Probe and Irrelevant conditions were statistically compared at group and individual level. As mentioned in subsection "Time-frequency analysis" of the "Methods" section, to obtain a single ITC measure for each condition, the sum of all positive changes in the ITC over 0.5 to 7 Hz frequencies was calculated. In this way, two single ITC measures were obtained for each participant, one for the Probe and the other for the Irrelevant. To compare these

measures at the group level, a two-tailed paired t-test was used. When applied to the orthogonal-contrast time windows, the outcome confirms highly significant differences between the ITC of the Probe and the Irrelevant ($p < 10^{-6}$) [6].

For individual level analysis, a randomization test was used with the ITC difference measures (described in subsection "Statistical Analysis of ERSP and ITC") to generate individual ITC null hypothesis distributions. When applied to an orthogonal-contrast time window, 93% (13/14) of the participants showed a significant difference between ITCs of Probe and Irrelevant. Ten participants had p-values below or equal to 0.001 and three participants obtained slightly higher p-values, but these were still below 0.05. Per-individual ITC p-values (calculated within an orthogonal-contrast time widow) are presented in Table 5.

Combined analysis (ERSP and ITC). Per-individual combined Fisher values of the ERSP and ITC transforms calculated within the 100 ms orthogonal-contrast time window are presented in Table 5. Using the Fisher combining procedure, we were able to combine the p-values of the ERSP and ITC for each participant into a single joint p-value. The combined p-values were smaller than the critical significance threshold (0.05) across all participants, resulting in a detection rate of 100% [7]. The average combined p-value of the 100 ms orthogonal-contrast time window was 0.000538.

Results of the Recognition Tests

As the experiment consisted of five blocks, each participant performed five recognition tests. All recognition tests were presented online at the end of each block. These tests were conducted to ensure that the Probe faces had indeed been recognized

[6] To clarify again, this result is obtained with notch filtering to remove the SSVEP set-up by the stream. Running the same analysis without notch filtering gives a p-value of 0.0109, i.e. the effect remains.
[7] As previously stated, this combined analysis was performed on data where we have notch filtered out the SSVEP. If we run the same analysis without the notch filter, 13 out of 14 participants have p-values below the critical alpha level.

by individuals. At the end of each block, Participants were instructed to give a confidence rating about how often a particular face was presented. Five categories of faces were presented: Target, Probe, Irrelevant, Un-presented Famous face, and a distractor. Note, as we have five Probes and five Irrelevant faces across the course of the entire experiment, there were also five un-presented Famous faces and five distractor faces. Apart from the Target (as there was a single Target face presented across the entire experiment), a different face from each category was presented at the end of each block.

At the end of each block, each face was assigned a confidence rating by each participant. To calculate the final confidence rating for each face, we took the mean rating across the five blocks. The mean confidence rating for all participants are presented in Table 6. Our main comparisons were Probe (famous faces) ($M = 3.61$, $SD = 0.80$, $Mdn = 3.8$) against un-presented famous faces ($M = 1.21$, $SD = 0.278$, $Mdn = 1.2$); and Irrelevant ($M = 1.95$, $SD = 0.633$, $Mdn = 1.9$) against the distractor face ($M = 1.34$, $SD = 0.439$, $Mdn = 1.2$). Pairwise comparisons made with Wilcoxon's signed-rank test revealed a highly significant difference between the Probe and un-presented famous face confidence rating, $Z = -3.295$, $p < 0.00013$. With regard to the comparison between the Irrelevant faces and the distractor faces, the statistical test showed a significant difference between the confidence rating assigned to each condition; $Z = -2.34$, $p = 0.0168$. However, although, some participants (especially Participants 1 and 6) have given higher confidence rates to the Irrelevant than to the distractors, their Irrelevant ERPs did not show apparent effects.

Discussion

Bowman et al. (2013, 2014) showed that presenting stimuli in RSVP provides an accurate and countermeasure-resistant means to demonstrate familiarity with an

individual's own-name. In particular, we showed that using this paradigm, a participant's name broke through into awareness, resulting in the generation of a P3 that could be used to identify individuals who are not disclosing their identity. This raises the possibility that such an RSVP-based method could be used as a general purpose concealed information test. We have named the resulting technique, the Fringe-P3 method. In this paper, the aim was to investigate whether faces can be used in an RSVP (sub)liminal search paradigm (Bowman et al., 2013, 2014) to infer recognition of familiar faces. Specifically, this paper assesses whether famous faces can break into conscious awareness when presented in RSVP and, importantly, whether these "breakthrough" events can be detected by EEG on a per-individual basis. To answer these questions, an ERP experiment was conducted; this followed a protocol similar to that used in Bowman et al. (2013), with the exception that there were no instructions to conceal identification of the Probe.

As the key comparison in EEG-based concealed information tests is between the Probe and Irrelevant conditions, brain responses to famous faces (Probes) and unfamiliar faces (Irrelevants) were statistically compared in the ERP domain as well as (in a more exploratory analysis) in the time-frequency domain.

Group-level, Time Domain: In terms of ERP data results in this paper, the Probe (famous faces) elicited ERP responses that were different to those elicited by the Irrelevant (unfamiliar faces). In particular, an enhanced negative deflection followed by small positivity over a time window from 300 ms to 700 ms post-stimulus was elicited only by the Probe and not by the Irrelevant. This finding replicates previous studies that investigated familiarity effects using famous faces (Bentin & Deouell, 2000; Eimer, 2000; Touryan et al., 2011a). As an example, Figure C.1 depicts ERPs elicited by familiar and unfamiliar faces at midline electrodes from Experiment 2 in (Bentin &

Deouell, 2000). Statistical analysis revealed that familiar non-politicians' faces elicited significantly more negative potentials than unfamiliar faces within a time window from 250 ms to 500 ms, which is in agreement with our findings. It is worth noting the ongoing oscillations elicited at Pz and Oz in (Bentin & Deouell, 2000), which are somewhat similar to the oscillatory patterns we obtained in our data (see Figure 2). However, the ERPs at Fz and Cz in the two studies do not appear to be similar. The reasons for these variations may be the stimulus presentation differences between the studies; after all, we were presenting stimuli at RSVP rates and specifically observing a "perceptual breakthrough" response.

Indeed, a key point is that we were presenting a rapid stream of stimuli. A striking property of the brain's processing of RSVP is how few stimuli in such streams leave a durable memory trace. For example, Potter showed that recognition memory for arbitrary stimuli presented in RSVP was low, while detection performance, i.e. searching for a pre-specified target, is high (Potter, 1976). This is consistent with the perspective inherent in our term (sub)liminal salience search (Bowman et al, 2013) that the brain is searching on the fringe of awareness when viewing RSVP, and only a small subset of stimuli "breakthrough into awareness". In the experiments reported here, it would typically be that the famous faces would break-through into awareness. Importantly, this is the first time ERP components and oscillatory correlates for the break-through into awareness (associated with a presentation mode such as RSVP) of famous faces has been reported.

Group-level, Frequency Domain: Time frequency analyses were performed because of the sustained oscillatory activity elicited by the Probe (famous faces) starting as early as 75 ms and finishing around 650 ms, which would not be reflected in other ERP measurements (e.g. peak-to-peak, mean amplitude or base-peak). Because of

this, time frequency analyses were used to characterize these oscillations by measuring their power (ERSP) and phase coherence (ITC) across replications. Indeed, several studies have shown that changes in power or inter-trial phase coherence are likely to contribute to different ERP components (Cheron et al., 2007; Fuentemilla, Marco-Pallarés, Münte, & Grau, 2008). For example, Miyakoshi et al. found that the P3 component in response to one's own face was associated with increased power (ERSP) in the theta frequency band (4 Hz–7 Hz) (Miyakoshi, Kanayama, Iidaka, & Ohira, 2010).

Additionally, the fact that large increases in ERSP and ITC were observed only in the Probe condition within a time window from 250 ms to 520 ms suggests that such changes in power and across trial coherence may have contributed to the generation of the N400f in the same time window for the Probe ERP.

Our greatest interest here is not with induced responses (power increases not accompanied by phase resets), since the high inter-trial coherence that we observe suggests that the high power we see is not induced in nature. Our main interest is to determine if we obtain higher statistical power by capturing in our analysis the ongoing oscillatory pattern we observe, as discussed further shortly.These results support the claim that oscillatory activity and some ERP components reflect similar cognitive processes (Fuentemilla et al., 2008; Makeig et al., 2004; Yordanova, Kolev, & Polich, 2001).

Notably, the significant ITC starts considerably earlier (around 50 ms post stimulus) than the significant ERSP (around 200 ms post stimulus), compare Figure 7 top and bottom rows. This suggests that the (drunken-Wavelet) oscillatory pattern we observe begins with a pure phase reset, which evolves into an increase in oscillatory amplitude. This increase in amplitude is likely to be most dramatically marked by the

large N400f deflection we observe in our ERPs. It is also notable that our statistical effects both at the group and individual level, were larger for the time-frequency than ERP analysis. This does justify our intuition that an ongoing (i.e. multi-cycle) evoked oscillatory pattern is present in our data.

Per-Individual, Time Domain: With regard to detection rate (i.e. the percentage of participants who elicited significantly stronger effects for the Probe than for the Irrelevant) 71% (10/14) of participants, elicited statistically significant negative deflection ('N400f') components in response to the Probe as compared to the Irrelevant. With regard to the positive deflection ('P600'), only seven (50% of) participants showed strong P600 Probe effects. In order to obtain a single measurement for each participant, the N400f was combined with the P600, using the Fisher combining technique. This resulted in a 100 % (14/14) detection rate and an average p-value of 0.0098.

Per-Individual, Frequency Domain: The high detection rates obtained from the analysis of ERP data were supported by results from time-frequency analysis. For the ERSP, there was increased power (0.5–10 Hz) in a time window from about 250 ms to 520 ms (post-stimulus onset) only in the Probe as compared to the Irrelevant trials. All participants had p-values less than 0.05, resulting in a detection rate of 100%. Similar findings were obtained for the ITC transform, which show that the Probe produced a high ITC in a time window between ~150 ms and 520 ms (post-stimulus onset) and over a frequency band of 0.5–10 Hz, whereas in the Irrelevant condition, there were no significant changes in the ITC within the same time-window and over the same frequencies. Per-individual ITC statistical analysis showed that the ITC was statistically higher in the Probe than in the Irrelevant; 93% (13/14) of the participants showed a significant difference between ITCs of Probe and Irrelevant. As in the ERP

analysis, Fisher's combining method was also used to combine the ERSP and ITC findings of each individual. This resulted in a 100% detection rate.

Broader Fringe-P3 Work: In relation to the larger Fringe-P3 strand of research, it is surprising that a clear P3 was not observed for the (famous face) Probes in this experiment. We have run these Fringe-P3 experiments with many different stimulus types: own-name, own-email-address, famous name and famous faces. In all these cases, we have observed P3b patterns, often combined with a P3a, apart from with famous faces. Why exactly this is the case remains unclear and awaits further studies, especially with dense electrode arrays and source localisation.

Fisher Method: Some might argue over the appropriateness of the Fisher combining method for the analysis used in our study, especially since it works perhaps surprisingly well with our data, meaning that one might think the method inflates the false positive rate (i.e. increases the probability of detecting an effect when none exists or, in other words, increases the Type I error rate). However, the Type I error rate of the Fisher method was assessed in Bowman et al. (2013) using artificial data sets in which we already know that the null hypothesis is true (i.e. save for sampling error, there was no difference between the conditions). The results confirm that the false-positive rate of the method is not larger than the alpha level (i.e. no inflation of the Type I error rate was found).

Moreover, in Bowman et al. (2013), we showed that when combining a pair of p-values, the Fisher combining method provides a much smaller combined p-value than the average of the pair in two cases. The first is when one of the p-values is highly significant, but the other is far from significance (i.e. much larger than the alpha level). This offers a 'disjunctive' element to combining, i.e. a bias towards the minimum of the

two p-values. This can be seen in our data; see Participants 6, 9 and 14 in Table 4. The second is if both p-values are almost significant (slightly higher than the alpha level): in this case, the Fisher method can yield a p-value that is smaller than the average of the two p-values. This offers a 'conjunctive' element to combining, i.e. a bias towards simultaneously low p-values. As an example of this, see Participant 5 in Table 4. For more explanations about these two cases, refer to the simulation study of the Fisher method in the Appendix, S1 and S2 of Bowman et al. (2013).

Conclusions: Results from the ERP and time-frequency data provide evidence that the Probe (famous faces) was differentially perceived and processed by participants' brains as compared to novel (or unfamiliar) faces (the Irrelevant). Although the Probe and Irrelevant were treated equally in the experiment, EEG data revealed large differences in brain response between these conditions. These differences support the hypothesis in (Bowman et al., 2013) that during RSVP, salient stimuli (e.g. famous faces) can reach consciousness and so generate pronounced electrical responses (as seen in the Probe data here), whereas novel stimuli (e.g. unfamiliar faces) are not perceived sufficiently to encode into working memory and should not, therefore, generate a differentiable electrical response. When placed within the context of (Bowman et al., 2013, 2014), the present results clarify the potential for using not only word-based stimuli but also faces in RSVP-based concealed information tests. Most significantly, the results (especially the detection rates in per-individual analysis) suggest that this approach could be used to determine whether a suspect has knowledge of a particular person's face. This would be particularly relevant to applications in which one is seeking to demonstrate that an individual has long-term familiarity with somebody else; this, for example, could be with a crime compatriot or with a member of a terrorist cell.

The work presented here is though still only an initial step towards a workable system. It is indicative that we have identified clear effects with famous face stimuli, but more experiments need to be run. The sort of familiarity we have for the faces of real-life acquaintances may be somewhat different to those we have for famous faces, which we may, for example, view en-face more than normal acquaintances. Thus, a next step is to run similar experiments to those presented here, but with images of personal friends.

Additionally, the experiments presented here may not be a perfect precedent for familiarity set-up by fleeting glimpses, e.g. of a mugger. To assess the suitability of RSVP for such an application, one could set-up an experiment in which participants see glimpses of otherwise unfamiliar faces and then familiarity for those faces is tested in RSVP presentations. Experiments of this kind in which random faces are presented to participants in controlled ways, enables full counter-balancing of stimuli, definitively ruling out influences of low-level visual features, something we have only been able to do using statistical comparisons of image statistics, which does not give as much assurance as full counter-balancing.

References

Bentin, S., & Deouell, L. Y. (2000). Structural encoding and identification in face processing: erp evidence for separate mechanisms. Cognitive Neuropsychology, 17(1), 35–55. doi:10.1080/026432900380472.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. The Behavioral and Brain Sciences, 30(5–6), 481–99; discussion 499. doi:10.1017/S0140525X07002786.

Bowman, H., Filetti, M., Alsufyani, A., Janssen, D., & Su, L. (2014). Countering countermeasures: Detecting identity lies by detecting conscious breakthrough. PLOS ONE, 9(3), e90595. doi:10.1371/journal.pone.0090595.

Bowman, H., Filetti, M., Janssen, D., Su, L., Alsufyani, A., & Wyble, B. (2013). Subliminal salience search illustrated: EEG identity and deception detection on the fringe of awareness. PLOS ONE, 8(1), e54258. doi:10.1371/journal.pone.0054258.

Brooks, J. L., Zoumpoulaki, A., & Bowman, H. (2017). Data-driven region-of-interest selection without inflating type I error rate. Psychophysiology, 54(1), 100–113. doi:10.1111/psyp.12682.

Bruce, V., & Young, A. (1986). Understanding face recognition. British Journal of Psychology, 77(3), 305–327. doi:10.1111/j.2044-8295.1986.tb02199.x.

Cheron, G., Cebolla, A. M., De Saedeleer, C., Bengoetxea, A., Leurs, F., Leroy, A., & Dan, B. (2007). Pure phase-locking of beta/gamma oscillation contributes to the N30 frontal component of somatosensory evoked potentials. BMC Neuroscience, 8(1), article 75.. doi:10.1186/1471-2202-8-75.

Curran, T., & Hancock, J. (2007). The FN400 indexes familiarity-based recognition of faces. NeuroImage, 36(2), 464–471. doi:10.1016/j.neuroimage.2006.12.016.

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. doi:10.1016/j.jneumeth.2003.10.009.

Dietrich, A. B., Hu, X., & Rosenfeld, J. P. (2014). The effects of sweep numbers per average and protocol type on the accuracy of the P300-based concealed information test. Applied Psychophysiology and Biofeedback, 39(1), 67–73. doi:10.1007/s10484-014-9244-y.

Eimer, M. (2000). Event-related brain potentials distinguish processing stages involved in face perception and recognition. Clinical Neurophysiology, 111(4), 694–705. doi:10.1016/S1388-2457(99)00285-0.

Friston, K. J., Rotshtein, P., Geng, J. J., Sterzer, P., & Henson, R. N. (2006). A critique of functional localisers. NeuroImage, 30(4), 1077–1087. doi:10.1016/j.neuroimage.2005.08.012.

Fuentemilla, L., Marco-Pallarés, J., Münte, T. F., & Grau, C. (2008). Theta EEG oscillatory activity and auditory change detection. Brain Research, 1220, 93–101. doi:10.1016/j.brainres.2007.07.079.

Ganis, G., & Patnaik, P. (2009). Detecting concealed knowledge using a novel attentional blink paradigm. Applied Psychophysiology and Biofeedback, 34(3), 189–196. doi:10.1007/s10484-009-9094-1.

Hayasaka, S., & Nichols, T. E. (2004). Combining voxel intensity and cluster extent with permutation test framework. NeuroImage, 23(1), 54–63. doi:10.1016/j.neuroimage.2004.04.035.

Henson, R. N., Goshen-Gottstein, Y., Ganel, T., Otten, L. J., Quayle, A., & Rugg, M. D. (2003). Electrophysiological and haemodynamic correlates of face

perception, recognition and priming. Cerebral Cortex, 13(7), 793–805. doi:10.1093/cercor/13.7.793.

Kaufmann, T., Schulz, S. M., Grünzinger, C., & Kübler, A. (2011). Flashing characters with famous faces improves ERP-based brain-computer interface performance. Journal of Neural Engineering, 8(5), 056016. doi:10.1088/1741-2560/8/5/056016.

Kilner, J. M. (2013). Bias in a common EEG and MEG statistical analysis and how to avoid it. *Clinical Neurophysiology : Official Journal of the International* Federation of Clinical Neurophysiology, 124(10), 2062–2063. doi:10.1016/j.clinph.2013.03.024.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009a). Circular analysis in systems neuroscience : The dangers of double dipping. Nature Neuroscience, 12(5), 535–540. doi:10.1038/nn.2303.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009b). Circular analysis in systems neuroscience: The dangers of double dipping. Nature Neuroscience, 12(5), 535–540. doi:10.1038/nn.2303.

Labkovsky, E., & Rosenfeld, J. P. (2012). The P300-based, complex trial protocol for concealed information detection resists any number of sequential countermeasures against up to five irrelevant stimuli. Applied Psychophysiology and Biofeedback, 37(1), 1–10. doi:10.1007/s10484-011-9171-0.

Lefebvre, C. D., Marchand, Y., Smith, S. M., & Connolly, J. F. (2007). Determining eyewitness identification accuracy using event-related brain potentials (ERPs). Psychophysiology, 44(6), 894–904. doi:10.1111/j.1469-8986.2007.00566.x.

Luck, S. J. (2005). An introduction to the event-related potential technique. Cambridge, MA: MIT Press. doi:10.1118/1.4736938.

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. Trends in Cognitive Sciences, 8(5), 204–210. doi:10.1016/j.tics.2004.03.008.

Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. Science, 295(5555), 690–694. doi:10.1126/science.1066168.

Meijer, E. H., Smulders, F. T. Y., Merckelbach, H. L. G. J., & Wolf, A. G. (2007). The P300 is sensitive to concealed face recognition. International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology, 66(3), 231–237. doi:10.1016/j.ijpsycho.2007.08.001.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, & Computers, 36(4), 630–633. doi:10.3758/BF03206543.

Ming, D., An, X., Xi, Y., Hu, Y., Wan, B., Qi, H., Cheng, L., & Xue, Z. (2010). Time-locked and phase-locked features of P300 event-related potentials (ERPs) for brain–computer interface speller. Biomedical Signal Processing and Control, 5(4), 243–251. doi:10.1016/j.bspc.2010.08.001.

Miyakoshi, M., Kanayama, N., Iidaka, T., & Ohira, H. (2010). EEG evidence of face-specific visual self-representation. NeuroImage, 50(4), 1666–1675. doi:10.1016/j.neuroimage.2010.01.030.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. Psychophysiology, 37(2), 127–152. doi:10.1111/1469-8986.3720127.

Potter, M. C. (1976). Short-term conceptual memory for pictures. Journal of experimental psychology: human learning and memory, 2(5), 509.

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. Psychophysiology, 41(2), 205–219. doi:10.1111/j.1469-8986.2004.00158.x.

Touryan, J., Gibson, L., Horne, J. H., & Weber, P. (2011a). Real-time measurement of face recognition in rapid serial visual presentation. Frontiers in Psychology.MAR, 2, 42. doi:10.3389/fpsyg.2011.00042.

Touryan, J., Gibson, L., Horne, J. H., & Weber, P. (2011b). Real-time measurement of face recognition in rapid serial visual presentation. Frontiers in Psychology, 2, 42. doi:10.3389/fpsyg.2011.00042.

Trenner, M. U., Schweinberger, S. R., Jentzsch, I., & Sommer, W. (2004). Face repetition effects in direct and indirect tasks: An event-related brain potentials study. Brain Research. Cognitive Brain Research, 21(3), 388–400. doi:10.1016/j.cogbrainres.2004.06.017.

Yordanova, J., Kolev, V., & Polich, J. (2001). P300 and alpha event-related desynchronization (ERD). Psychophysiology, 38(1), 143–152. doi:10.1111/1469-8986.3810143.

Author Notes

Correspondence should be addressed to Abdulmajeed Alsufyani,

P.O.Box 888, 21974. Al-Haweiah, Taif, KSA.

Tel: 00966506717711

Email: a.s.alsufyani@tu.edu.sa

Tables

Table 1: Outcomes of the statistical analysis of pixel intensity (first four moments) of the images of famous (Probes) and unfamiliar (Irrelevant) faces.

Moment	Probe (famous faces)	Irrelevant (unfamiliar faces)	Two-samples independent t-test
Mean	*M*=158.33, *SD*=12.02	*M*=166.81, *SD*=12.30	*t*=1.55, *p*=0.134
Variance	*M*=3.04e+03, *SD*= 1.2e+03	*M*=2.54e+03, *SD*=1.0e+03	*t*=-0.96, *p*=0.349
Skewness	*M*=-1.06, *SD*=0.526	*M*=-1.47, *SD*=0.322	*t*=2.07, *p*=0.052[8]
Kurtosis	*M*=4.21, *SD*=2.06	*M*=3.16, *SD*=1.16	*t*=-1.4, *p*=0.178

[8] The skewness value of one of the famous faces used in the recognition test was more negative than the average skewness values of all other faces (in famous and unfamiliar categories). As a result of this, the p-value of the skewness was almost significant, p = 0.052. We re-ran the analysis after excluding this face and the p-value of the skewness was 0.117. Importantly, since this face did not appear in the ERP part of the experiment, as it was only used as a face in the recognition test, it could not contribute to our ERP findings.

Table 2: Per-individual mean amplitude values of N400f and P600 in the Probe and Irrelevant conditions at Pz.

Participant	N400f		P600	
	Probe	Irrelevant	Probe	Irrelevant
1	-9.1349	-1.3518	3.7236	-1.7585
2	-7.7482	1.0752	4.0965	1.3300
3	-6.7583	0.2432	3.2797	-1.5746
4	-3.1936	1.7057	4.3932	1.3179
5	-4.1186	-0.9497	2.2443	-0.9545
6	-4.1316	1.3559	1.7909	1.4321
7	-1.9274	-0.6029	2.5197	-1.0975
8	-5.5501	-1.3248	-0.6807	0.4171
9	-4.0365	0.8894	1.1608	0.5636
10	-3.5774	-0.7039	3.6935	1.2010
11	-3.1664	-2.8961	2.1609	-1.7521
12	-4.0082	-0.7272	2.5482	-0.8100
13	-1.9829	0.6339	4.8537	-0.0651
14	-2.2547	1.1639	0.9091	-0.6671

This table presents per-individual mean amplitude values of the N400f and P600. The size (mean amplitude) of the N400f, respectively P600, for each condition (Probe and Irrelevant) was quantified by finding the N400f, respectively P600, time window from the aggregated grand average of trials (AGAT) of all Probe and Irrelevant trials across all participants.

Table 3 Summary of the group level-analyses of Probe N400f/P600 and Irrelevant N400f/P600 across the mid-line electrodes.

Channel	N400f			P600		
	Difference of mean amplitude	t-value	p-value	Difference of mean amplitude	t-value	p-value
Fz	-2.5347	-4.7560	0.00037	2.7573	3.2376	0.0065
Cz	-3.7987	-6.1377	0.000035	2.5203	3.4301	0.0045
Pz	-4.3992	-6.6823	0.000015	2.6210	5.5597	0.0000923

Table 4 Per-individual p-values and Fisher combined probability values.

Participants	N400f	P600	Fisher
1	0.0040	0.0480	<0.001
2	<0.001	0.0860	<0.001
3	<0.001	0.0020	<0.001
4	<0.001	0.0430	<0.001
5	0.0600	0.0830	0.0130
6	0.0110	0.3650	0.0140
7	0.1040	<0.001	<0.001
8	0.0170	0.4650	0.0300
9	0.0190	0.4960	0.0330
10	0.0040	0.0010	<0.001
11	0.4530	0.0170	0.0280
12	0.0120	0.2890	0.0140
13	0.1100	<0.001	<0.001
14	0.0010	0.4910	<0.001

This table presents p-values of the N400f and P600 obtained from the randomization test with mean amplitude measures and the Fisher combining procedure, for all participants. As can be seen, Fisher p-values, which we used to determine whether the ERP response elicited by the Probes (famous faces) was significantly different from that elicited by the Irrelevant (unfamiliar faces), are all below 0.05, resulting in a 100% detection rate.

Table 5 Per-individual ERSP and ITC p-values and Fisher combined probability value, *calculated using a participant's orthogonal*-contrast time window selected from the aggregated power/coherence.

Participants	ERSP	ITC	Fisher
1	0.0001	0.0001	<0.0001
2	0.0001	0.0094	<0.0001
3	0.0001	0.0001	<0.0001
4	0.0142	0.6791	0.0062
5	0.0001	0.0001	<0.0001
6	0.0025	0.0001	<0.0001
7	0.0073	0.0001	<0.0001
8	0.0007	0.0002	<0.0001
9	0.0245	0.0008	<0.0001
10	0.0001	0.0001	<0.0001
11	0.0130	0.0004	<0.0001
12	0.0001	0.0018	<0.0001
13	0.0001	0.0001	<0.0001
14	0.0007	0.0155	0.0001

Table 6 Confidence ratings obtained from the recognition test

Participant	Probe	Irrelevant	Un-presented famous face	Distractor
1	3.8	3	1	1.2
2	4.4	1	1	1
3	4.2	1.6	1.2	1.4
4	4.4	1.8	1.4	1
5	3.2	1.2	1.2	2.6
6	4	3	1	1.8
7	3.2	2.2	2	1.4
8	2.2	2	1.2	1
9	4.2	2.6	1.6	1.6
10	3.4	1.6	1	1.2
11	3.8	2	1.2	1.4
12	4.4	1.8	1	1
13	3.6	2.4	1	1.2
14	1.8	1.2	1.2	1
Average	**3.61**	**1.95**	**1.21**	**1.34**

Shown above are the confidence ratings obtained from the recognition tests. At the end of each block, participants were presented with all faces from the five categories and asked to give a confidence rating about how often they appeared[9]. They were asked to do so with a number ranging from 1 to 5, where 5 indicated that they saw the face frequently and 1 indicated that they did not see the face. As there were five blocks (i.e.

[9] The confidence ratings of the fifth category, which was the Target face, is not presented here, as the participants were informed about it and instructed to respond to it at the end of each trial.

five recognition tests) in the experiment, the final rating for each condition was calculated by taking the mean of the five ratings.

Figure legends

Figure 1: Part of an RSVP stream of face stimuli.

Each RSVP stream consisted of 17 random (unfamiliar) faces used as distractors (or fillers), and only one critical face (could be Probe, Irrelevant, or Target); a Probe (i.e. famous) face is presented in this figure as a critical face. The SOA was 133 ms.

Figure 2: Grand average ERPs of all conditions at Fz, Cz and Pz.

Grand average ERPs elicited by the Target, Probe (famous faces), and Irrelevant (unfamiliar faces) at midline electrodes (Fz, Cz and Pz). As the Target was made task-relevant by instruction, it elicited a large P3. The key comparison is between the Probe and Irrelevant ERPs. As can be seen at each of the midline electrodes, a multi-cycle oscillation pattern (with a frequency of ~3.5-4 Hz) starting at around 100 ms and finishing around 650 ms was elicited only by the Probe. From within this oscillatory pattern, in a time window from 300 to 620 ms the Probe elicited an enhanced negativity (which we interpret as an N400f) followed by a positive deflection (which we interpret as a P600), whereas the Irrelevant showed no evidence of such negativity or positivity.

Figure 3: Individuals' Probe and Irrelevant ERPs at Pz.

Positive is plotted upwards. Each ERP is labelled with the corresponding participant number. The thinner line represents the Irrelevant ERP, while the thicker line represents the Probe ERP. As can be observed, the Probe of most of the participants elicited a distinct pattern (negative deflections followed by positive) within a time window from about 300 ms to 700 ms with respect to the stimulus onset. Note the absence of such a pattern in the Irrelevant ERPs. Moreover, for most participants, the N400f and P600 deflections can be placed within a longer oscillatory pattern that may start as early as 75 ms post Probe onset.

Figure 4: N400f Probe minus N400f Irrelevant mean amplitude null hypothesis distributions for Participants 2 and 11.

The dashed vertical lines represent the true observed value (N400f Probe – N400f Irrelevant mean amplitude). The true observed value for Participant 2, who has a strong effect, falls outside the null hypothesis distribution, resulting in a highly significant p-value ($p < 0.001$). On the other hand, the null hypothesis distribution for Participant 11 (who elicited the weakest effect) shows that the true observed value was too small to reject the null hypothesis, which led to a high p-value: $p = 0.453$.

Figure 5: Fisher values distribution for Participant 9.

Fisher (N400f and P600 combined) null hypothesis distribution for Participant 9 (who has the highest combined p-value: $p = 0.033$). The dashed vertical line represents the true observed Fisher value and p-value region.

Figure 6: ERSP and ITC results of the aggregated grand average of trials (AGAT) of the Probe and irrelevant conditions combined together (across 14 participants) at Pz.

Figure 6 presents two time-frequency plots. The ERSP plot of the AGAT is presented in the top panel, while the ITC plot of the AGAT is presented in the lower panel. Each transform was calculated by pooling all trials from both Irrelevant and Probe conditions across all participants. In each plot, the vertical purple dashed line marks the onset of the critical (Probe/Irrelevant) stimulus presentation, whereas the two black dashed vertical lines indicate the start and the end of the time window selected under the AGAT (orthogonal contrast) for the group level analyses. Note that this figure was produced after applying a 7 – 9 Hz notch filter, so that the SSVEP elicited by the stream presentation frequency had been filtered out.

Figure 7: ERSP & ITC results of the global trial-by-trial (across all participants) analysis for the Probe and Irrelevant conditions at Pz.

Figure 7 presents six time-frequency plots: upper row are ERSP plots and lower ITC. The plots of the Probe are presented in the first column, with the Irrelevant plots in the second and the difference between Probe and Irrelevant plots in the third. For each

condition, ERSP and ITC were calculated by analyzing all trials from 14 participants. On the right of each plot, a color bar shows the ERSP and ITC scales. An increase in power and ITC (relative to the baseline period) is colored by red, whereas a decrease is indicted by blue. Green areas in each plot indicate no significant changes. In each plot, the vertical purple dashed line marks the onset of the critical (Probe/Irrelevant) stimulus presentation. With regard to ERSP (top row), compared with the Irrelevant, the Probe elicited significantly higher power in a time window from about 280 to 520 ms and over 0.5-10Hz frequencies. The difference between the ERSPs of the two conditions within the same time window and frequencies is shown in the (Probe minus Irrelevant) ERSP plot, top right panel. With regard to ITC (bottom row), as can be seen from the ITC plot of the Probe and Irrelevant, high ITC values were produced in a time window from about 150 ms to 520 ms and over a 0.5-10 Hz-frequency band for the Probe trials, but not with the Irrelevant. The difference between the ITC of the two conditions within the same time window and frequencies is shown in the (Probe minus Irrelevant) ITC plot, bottom rightmost panel.

Figure A.1: Stimuli of Famous (Probe) and unfamiliar (Irrelevant) faces.

Figure A.1 presents examples of faces stimuli used in the experiment. The first row contains five famous faces used as Probe stimuli, while the second row shows five unfamiliar stimuli used as Irrelevants.

Figure B.1: ERSP and ITC results of (AGAT) in which the SSVEP set up by the stream presentation was not filtered out .

Figure B.1 presents two time-frequency plots. The ERSP plot of the AGAT is presented in the top panel, while the ITC plot of the AGAT is presented in the lower panel. Each transform was calculated by pooling all trials from both Irrelevant and Probe conditions across all participants. In each plot, the vertical purple dashed line marks the onset of the critical (Probe/Irrelevant) stimulus presentation, whereas the two black dashed vertical lines indicate the start and the end of the orthogonal-contrast time window in which the ERSP/ITC in both conditions were measured for the group level analyses**.**

Note that filtering out the SSVEP generated by the stream presentation frequency (by applying a notch filter between 7-9 Hz) produces a cleaner ITC pattern (presented in Figure 6).

Figure C.1: Grand average ERPs of familiar and unfamiliar faces at midline electrodes from experiment 2 in (Bentin & Deouell, 2000).

The above figure presents ERPs elicited by familiar and unfamiliar faces. In this experiment, participants were presented with 180 faces (60 were of famous politicians, 60 of famous people from different fields apart from politics, and 60 of unfamiliar people). They were instructed to count the number of times that faces of politicians were presented. The main comparison was between ERPs of familiar non-politicians and unfamiliar faces. Statistical analysis of the ERPs revealed that familiar non-politicians faces elicited significantly more negative potentials than those elicited by unfamiliar faces within a time window from 250 ms to 500 ms.

Figures

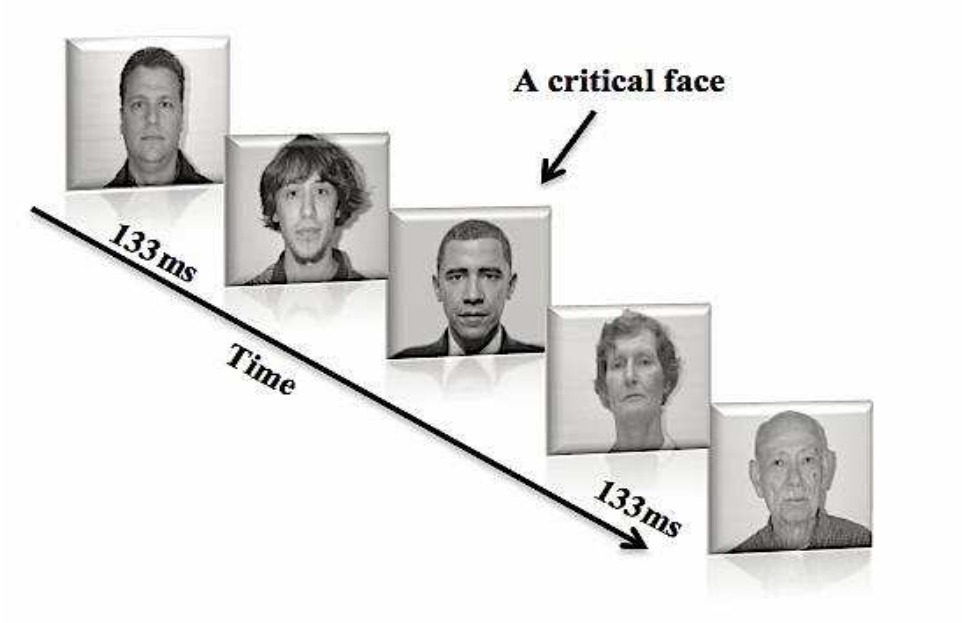

Figure 1: Part of an RSVP stream of face stimuli.

Each RSVP stream consisted of 17 random (unfamiliar) faces used as distractors (or fillers), and only one critical face (could be Probe, Irrelevant, or Target); a Probe (i.e. famous) face is presented in this figure as a critical face. The SOA was 133 ms.

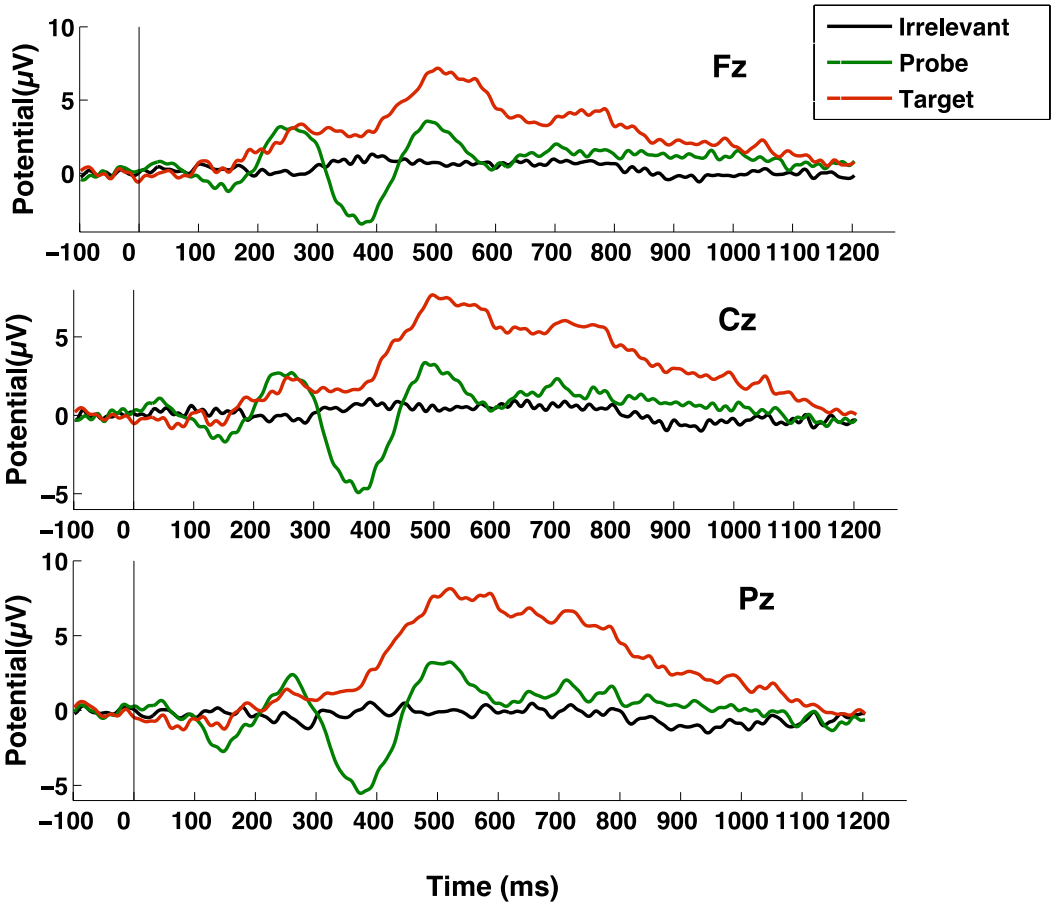

Figure 2: Grand average ERPs of all conditions at Fz, Cz and Pz.

Grand average ERPs elicited by the Target, Probe (famous faces), and Irrelevant (unfamiliar faces) at midline electrodes (Fz, Cz and Pz). As the Target was made task-relevant by instruction, it elicited a large P3. The key comparison is between the Probe and Irrelevant ERPs. As can be seen at each of the midline electrodes, a multi-cycle oscillation pattern (with a frequency of ~3.5-4 Hz) starting at around 100 ms and finishing around 650 ms was elicited only by the Probe. From within this oscillatory pattern, in a time window from 300 to 620 ms the Probe elicited an enhanced negativity (which we interpret as an N400f) followed by a positive deflection (which we interpret as a P600), whereas the Irrelevant showed no evidence of such negativity or positivity.

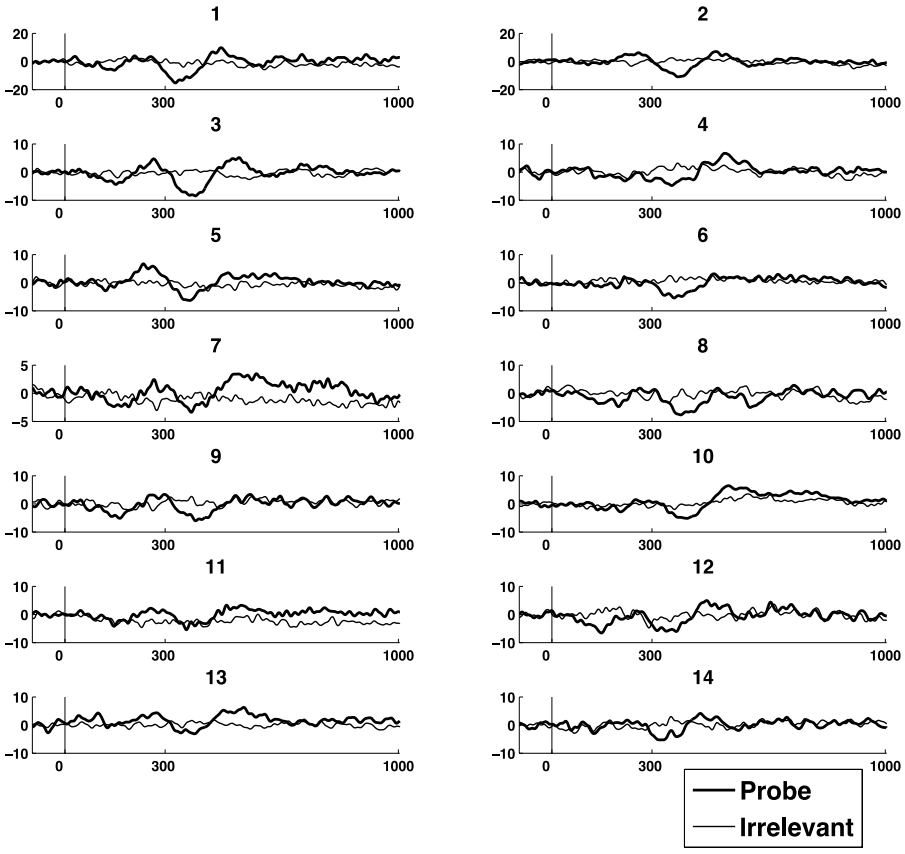

Figure 3: Individuals' Probe and Irrelevant ERPs at Pz.

Positive is plotted upwards. Each ERP is labeled with the corresponding participant number. The thinner line represents the Irrelevant ERP, while the thicker line represents the Probe ERP. As can be observed, the Probe of most of the participants elicited a distinct pattern (negative deflections followed by positive) within a time window from about 300 ms to 700 ms with respect to the stimulus onset. Note the absence of such a pattern in the Irrelevant ERPs. Moreover, for most participants, the N400f and P600 deflections can be placed within a longer oscillatory pattern that may start as early as 75 ms post Probe onset.

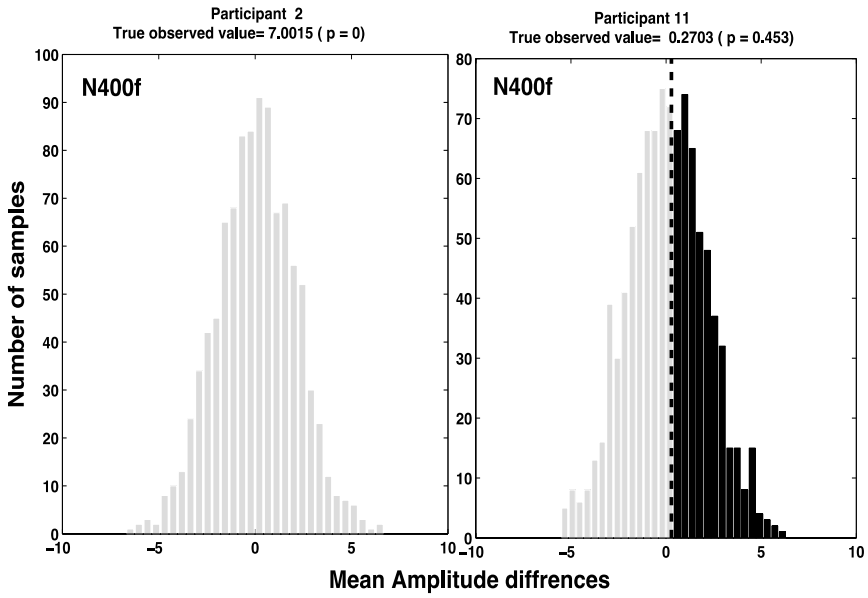

Figure 4: N400f Probe minus N400f Irrelevant mean amplitude null hypothesis distributions for Participants 2 and 11.

The dashed vertical lines represent the true observed value (N400f Probe – N400f Irrelevant mean amplitude). The true observed value for Participant 2, who has a strong effect, falls outside the null hypothesis distribution, resulting in a highly significant p-value ($p < 0.001$). On the other hand, the null hypothesis distribution for Participant 11 (who elicited the weakest effect) shows that the true observed value was too small to reject the null hypothesis, which led to a high p-value: $p = 0.453$.

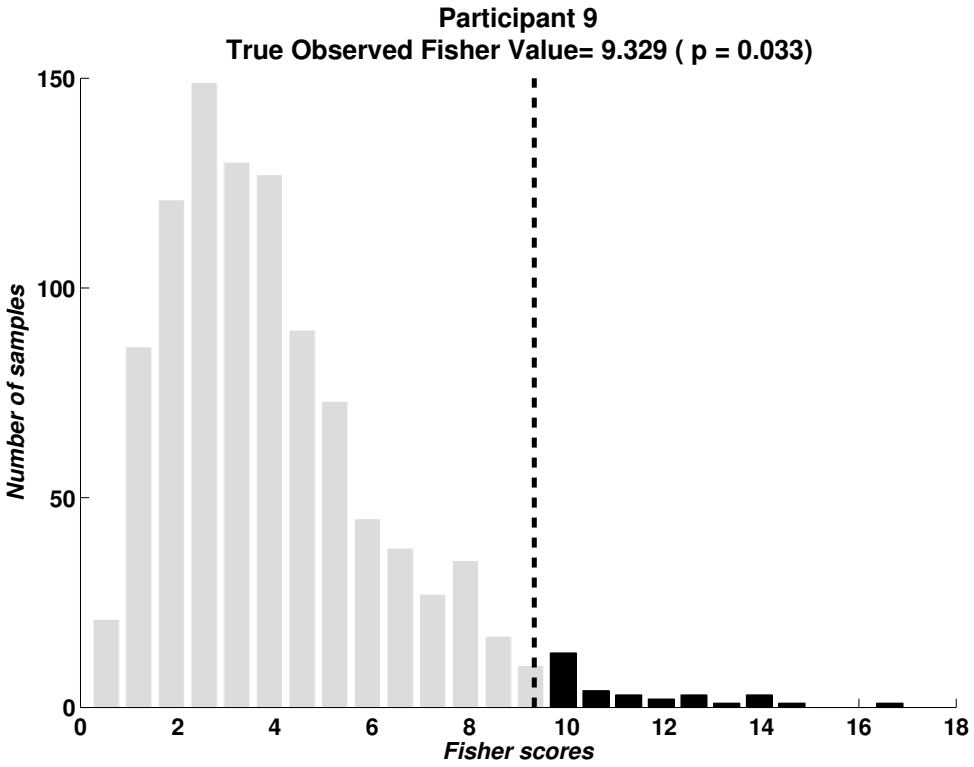

Figure 5: Fisher values distribution for Participant 9.

Fisher (N400f and P600 combined) null hypothesis distribution for Participant 9 (who has the highest combined p-value: p = 0.033). The dashed vertical line represents the true observed Fisher value and p-value region.

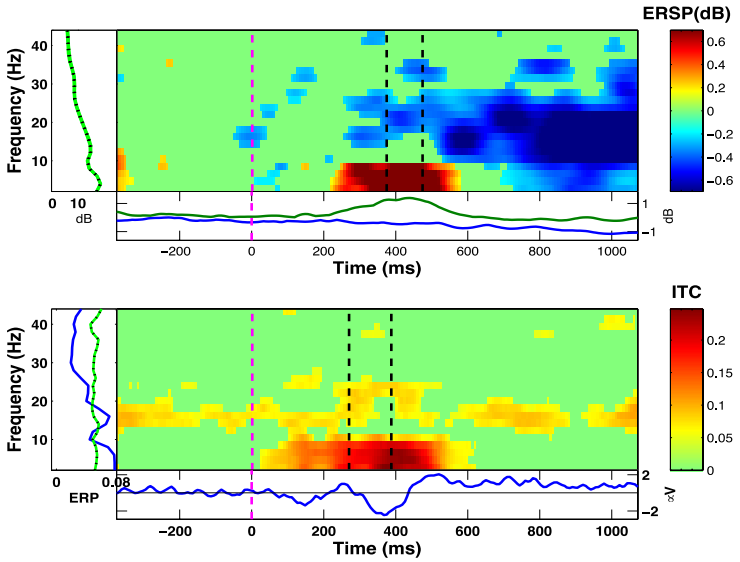

Figure 6 ERSP and ITC results of the aggregated grand average of trials (AGAT) of the Probe and irrelevant conditions combined together (across 14 participants) at Pz.

Figure 6 presents two time-frequency plots. The ERSP plot of the AGAT is presented in the top panel, while the ITC plot of the AGAT is presented in the lower panel. Each transform was calculated by pooling all trials from both Irrelevant and Probe conditions across all participants. In each plot, the vertical purple dashed line marks the onset of the critical (Probe/Irrelevant) stimulus presentation, whereas the two black dashed vertical lines indicate the start and the end of the time window selected under the AGAT (orthogonal contrast) for the group level analyses. Note that this figure was produced after applying a 7 – 9 Hz notch filter, so that the SSVEP elicited by the stream presentation frequency had been filtered out. **The lower part of the top panel shows two lines, indicating the ERSP envelope: low (in blue) and high (in green) mean dB vaule, relative to baseline.**

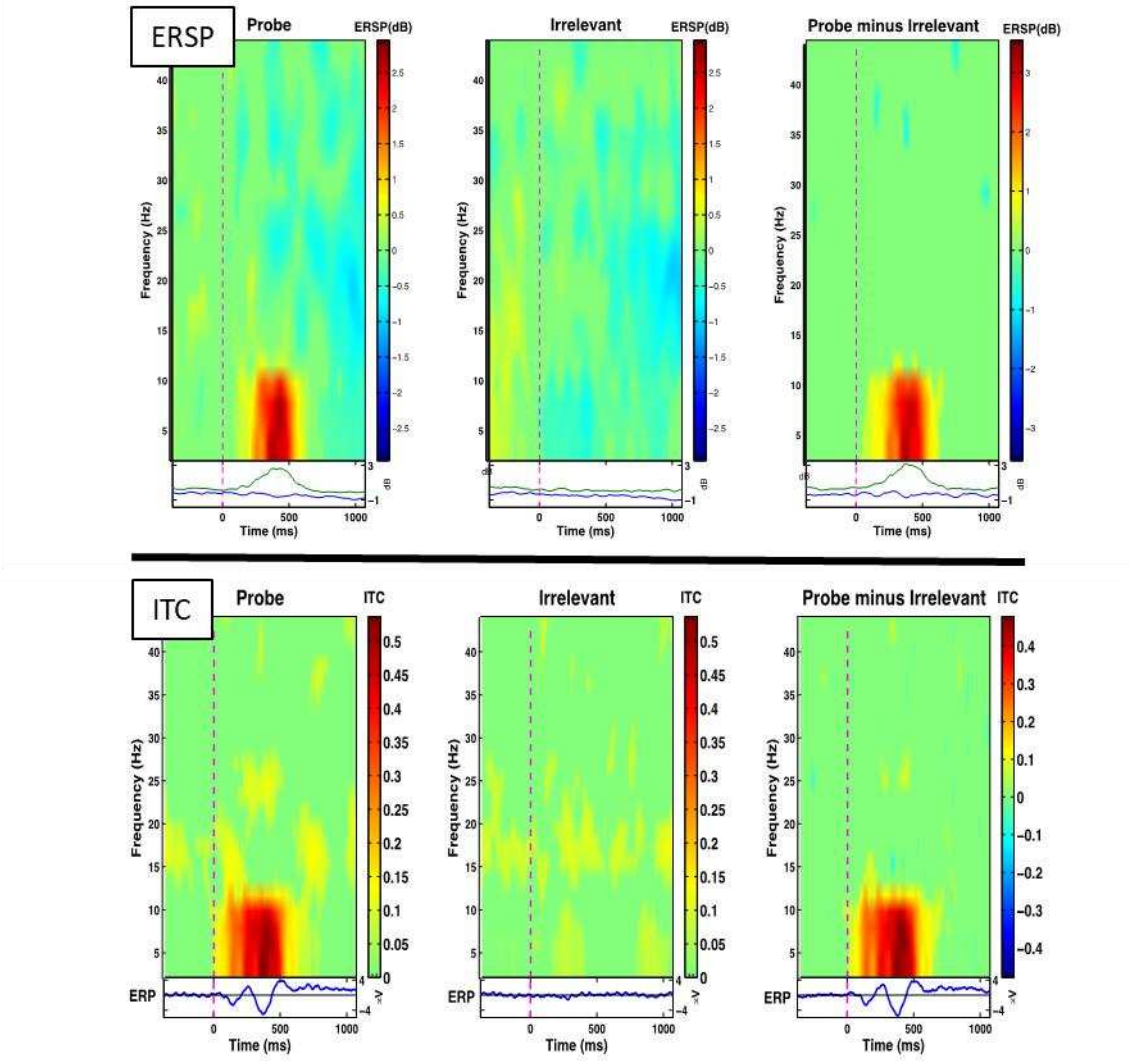

Figure 7: ERSP & ITC results of the global trial-by-trial (across all participants) analysis for the Probe and Irrelevant conditions at Pz.

Figure 7 presents six time-frequency plots: upper row are ERSP plots and lower ITC. The plots of the Probe are presented in the first column, with the Irrelevant plots in the second and difference between Probe and Irrelevant plots in the third. For each condition, ERSP and ITC were calculated by analyzing all trials from 14 participants. On the right of each plot, a color bar shows the ERSP and ITC scales. An increase in

power and ITC (relative to the baseline period) is colored by red, whereas a decrease is indicated by blue. Green areas in each plot indicate no significant changes. In each plot, the vertical purple dashed line marks the onset of the critical (Probe/Irrelevant) stimulus presentation. With regard to ERSP (top row), compared with the Irrelevant, the Probe elicited significantly higher power in a time window from about 280 to 520 ms and over 0.5-10Hz frequencies. The difference between the ERSPs of the two conditions within the same time window and frequencies is shown in the (Probe minus Irrelevant) ERSP plot, top right panel. With regard to ITC (bottom row), as can be seen from the ITC plot of the Probe and Irrelevant, high ITC values were produced in a time window from about 150 ms to 520 ms and over a 0.5-10 Hz-frequency band for the Probe trials, but not with the Irrelevant. The difference between the ITC of the two conditions within the same time window and frequencies is shown in the (Probe minus Irrelevant) ITC plot, bottom rightmost panel.

Appendix A

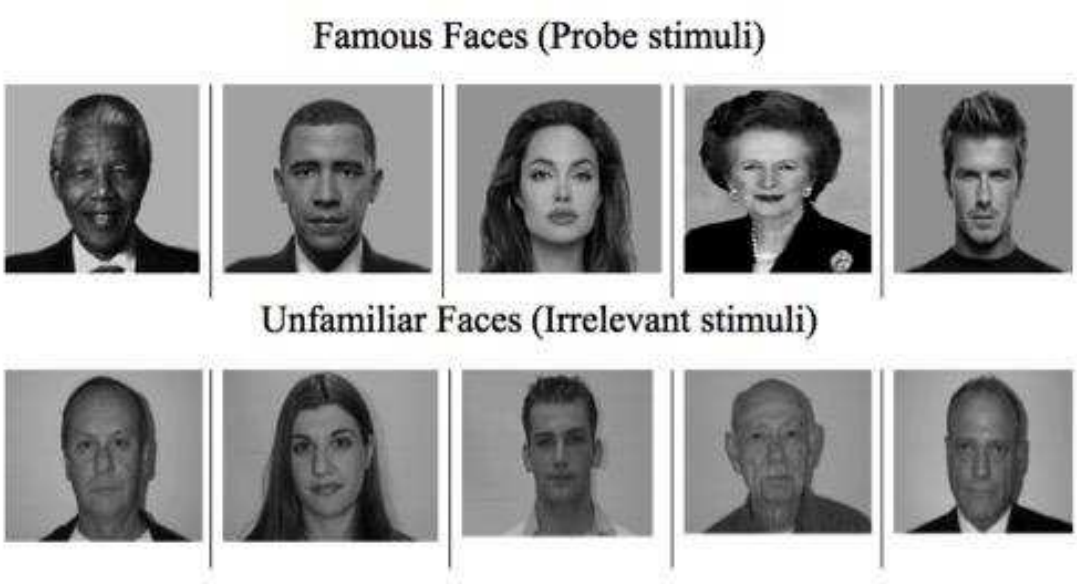

Figure A.1 Stimuli of Famous (Probe) and unfamiliar (Irrelevant) faces.

Figure A.1 presents examples of faces stimuli used in the experiment. The first row contains five famous faces used as Probe stimuli, while the second row shows five unfamiliar stimuli used as Irrelevant.

Appendix B

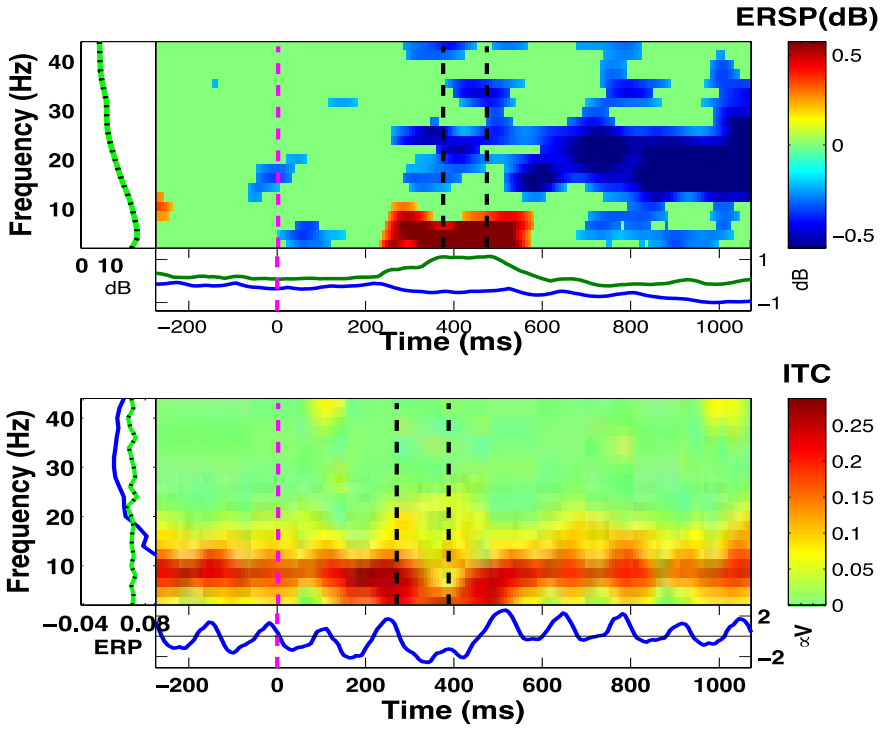

Figure B.1 ERSP and ITC results of (AGAT) in which the SSVEP set up by the stream presentation was not filtered out .

Figure B.1 presents two time-frequency plots. The ERSP plot of the AGAT is presented in the top panel, while the ITC plot of the AGAT is presented in the lower panel. Each transform was calculated by pooling all trials from both Irrelevant and Probe conditions across all participants. In each plot, the vertical purple dashed line marks the onset of the critical (Probe/Irrelevant) stimulus presentation, whereas the two black dashed vertical lines indicate the start and the end of the orthogonal-contrast time window in which the ERSP/ITC in both conditions were measured for the group level analyses**. The lower part of the top panel shows two lines, indicating the ERSP envelope: low (in blue) and high (in green) mean dB vaule, relative to baseline.**

As we did in Bowman et al (2014), we have notch filtered out the Steady State Visual Evoked Potential set-up by the presentation of the stimulus stream. This is in part because its presence confounds our time-frequency decomposition. One can see this by

comparing the time-frequency plots (Event Related Spectral Perturbation (ERSP) and Inter-trial Coherence (ITC)) for AGAT with notch filter (figure 6) and AGAT without notch filter (figure B.1). Specifically, the following are the key points with regard to this issue.

1) The ERSP plot for the AGAT without notch filter (figure B.1) shows that there is no differential power change at the frequency of the stream associated with presentation of critical stimuli. In fact, there is no above threshold ERSP signal across the entire time segment for the stream frequency, i.e. 7.5hz. If critical stimuli had generated an evoked or induced change in power, this would have been visible as an increase (relative to baseline) in ERSP at some point from zero onwards in the analysed segment. The fact that no such transient response is apparent suggests that a constant SSVEP is all that is present in the baseline period and analysis segment at the frequency of the stream.

2) Indeed, the ERSP plot suggests that there are two dominant evoked responses, neither of which is focused on the frequency of the stream. The first evoked pattern is the power increase (i.e. synchronization) at lower frequencies than the SSVEP. One might describe this as a theta burst, which is between 250ms and 550ms and corresponds to the drunken-wavelet. The second response is a higher frequency power decrease from 550ms out to the end of the analysed time segment. This though is present in both Irrelevant and Probe conditions, and thus does not allow us to distinguish famous and unfamiliar faces. Consequently, we do not consider it further.

3) The Inter-trial Coherence plot for the AGAT without notch filtering (figure B.1), though, shows a clear band of increased coherence across replications at the frequency

of the stream (7.5hz) (and also at its first harmonic). The presence of this band affects the ITC pattern corresponding to the theta burst between 250 and 550ms, due to inaccuracies in frequency decomposition. Such inaccuracies could only be resolved by sacrificing time-resolution, which we do not wish to do.

One can see the consequence of this band being present by comparing the ITC plot for the AGAT without notch filtering (figure B.1) and the AGAT with notch filtering (figure 6). The across trial phase synchrony associated with the drunken-wavelet found in the ERP is much more consistently present in the latter of these. In order to obtain a robust assessment of the ITC associated with the theta burst, we notch filtered out the SSVEP set-up by the RSVP stream.

Appendix C

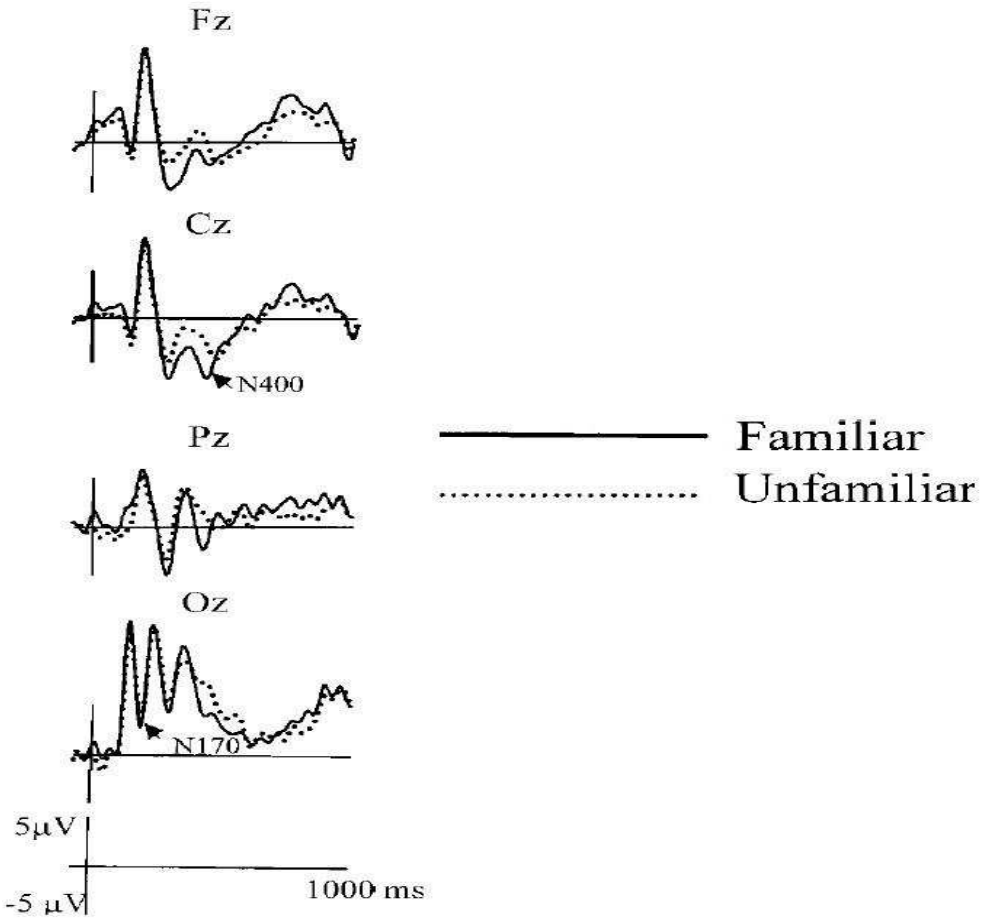

Figure C.1: Grand average ERPs of familiar and unfamiliar faces at midline electrodes from experiment 2 in (Bentin & Deouell, 2000).

The above figure presents ERPs elicited by familiar and unfamiliar faces. In this experiment, participants were presented with 180 faces (60 were of famous politicians, 60 of famous people from different fields apart from politics, and 60 of unfamiliar people). They were instructed to count the number of times that faces of politicians were presented. The main comparison was between ERPs of familiar non-politicians and unfamiliar faces. Statistical analysis of the ERPs revealed that familiar non-

politicians faces elicited significantly more negative potentials than those elicited by unfamiliar faces within a time window from 250 ms to 500 ms.