High test–retest reliability of checkerboard reversal visual evoked potentials (VEP) over 8 months

Johannes Sarnthein\textsuperscript{a,b,*}, Maria Andersson\textsuperscript{c}, Michael B. Zimmermann\textsuperscript{c}, Dominik Zumsteg\textsuperscript{d}

\textsuperscript{a}Neurochirurgische Klinik, UniversitätsSpital Zürich, Switzerland
\textsuperscript{b}Zürich Center for Integrative Human Physiology, Universität Zürich, Switzerland
\textsuperscript{c}The Human Nutrition Laboratory, Institute of Food Science and Nutrition, Swiss Federal Institute of Technology Zurich, Zurich, Switzerland
\textsuperscript{d}Neurologische Klinik, UniversitätsSpital Zürich, Switzerland

\section{Introduction}

Visual evoked potentials (VEP) refer to the cortical response recorded primarily from the occipital cortex following a standardized visual stimulation. Conventionally, VEP waveforms evoked by a reversing checkerboard stimulus are characterized by a negative deflection at 75 ms after the stimulus (N75) followed by a positive deflection at 100 ms (P100) (Halliday, 1982; Odom et al., 2004). The checkerboard reversal VEP is a standard tool in clinical research used to specifically assess the functioning of the central nervous system (CNS) (Brigell et al., 1994). For specific clinical questions, e.g. alteration in myelination of the optic nerve, the intra-individual stability is important for the follow-up of individual subjects.

The intra-individual stability of the VEP has been addressed in some publications, which are generally quite old (Hall et al., 1973; Oken et al., 1987; Schellberg et al., 1987; Joost et al., 1992). VEP amplitude is usually measured as the difference between peaks, but a frequency domain measure has also been proposed (Joost et al., 1992). To describe similarity in waveform, in these publications correlations between repeated measurements were calculated. However, authors agree that their approach to intra-individual VEP stability is not satisfactory (Oken et al., 1987; Schellberg et al., 1987; Joost et al., 1992).

As a new approach in VEP research, we compared test–retest VEPs by regressing VEP waveforms pairwise onto each other and calculated a t-value. This approach is similar to our earlier work, where we have analysed the test–retest reliability of EEG power spectra while subjects were resting with eyes closed (Näpflin et al., 2007) or performing a working-memory task (Näpflin et al., 2008). While the earlier analyses were in the frequency domain, we report here an application of the method to time-domain data. For repeated VEP measurements of individual subjects, the intra-individual stability is set in relation with the inter-individual

\textsuperscript{*}Corresponding author. Address: Neurochirurgische Klinik, UniversitätsSpital Zürich, Frauenklinikstrasse 10, CH-8091 Zürich, Switzerland. Tel.: +41 44 255 3672; fax: +41 44 255 4399.
E-mail address: johannes.sarnthein@usz.ch (J. Sarnthein).

1388-2457/$36.00 © 2009 International Federation of Clinical Neurophysiology. Published by Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.clinph.2009.08.014
variation of VEP waveforms. To characterize intra-individual stability of VEP waveforms over recording sessions, we investigated differences in the N75 latency, the P100 latency, the peak-to-peak amplitude and the shape of the VEP waveforms. We were interested how these observables would contribute to satisfactory recognition probabilities and thereby establish a statistical signature of persons on the basis of their VEP waveforms.

2. Methods

2.1. Study subjects

The VEP study was conducted in 10 female students 19–29 years old (median 22.4 years). At baseline, all subjects were screened for health problems using a brief medical history. Women were apparently healthy (except for iron deficiency), not pregnant or breast-feeding, and not using medication (except hormonal contraceptives). The subjects had no current or previous history of epilepsy. Vision of subjects was normal or corrected to normal with glasses or contact lenses. The study was approved by the Kanton Zürich ethics committee, Switzerland. All subjects were informed about the study and gave written informed consent according to the declaration of Helsinki.

2.2. VEP recording sessions

Subjects participated in two VEP sessions with a retest interval of 8 months. The recording sessions were performed between 17:00 h to 20:00 h in order to exclude an impact of circadian factors on the VEP. No coffee was consumed 3 h before the measurement to avoid effects of caffeine on the EEG (Landolt et al., 2004). For recording, subjects were seated in a dimly lit room shielded against sound and stray electric fields and were instructed to assume a comfortable position in a chair and place their head on a chin-rest. A computer screen displayed a black and white full-screen checkerboard with 16 black checks to the right eye. Mean luminance of the stimulus was 100 cd/m² with 98% contrast. The checkerboard was reversed in contrast at the rate of 2 reversals per second. At the viewing distance of 1.4 m the check edges subtended 30° of visual angle (spatial frequency of 1 cycle per degree).

EEG signals were measured with surface cup electrodes at the standard position Oz of the international 10–20 system with linked earlobes as reference. Impedances were below 10 kΩ. EEG signals were registered using the Neuropack 8 MEB-4200 K system (Nihon Kohden, www.nihonkohden.com, A/D conversion 12 bit, sampling rate 3500 Hz, band pass filter 0.5 Hz–200 Hz, -12 dB/octave) and continuously viewed on PC monitor. The responses to 100 checkerboard inversions were averaged to obtain one VEP waveform. The procedure was repeated three times and the three VEP waveforms averaged. All measurements were performed by the same investigator (M.A.). Data were stored in binary format on disk. For offline analysis, data were read into Matlab (www.TheMathworks.com) and processed using custom scripts. Before further processing, VEP waveforms were low-pass filtered digitally at 150 Hz.

2.3. Latency of N75 and P100 and peak-to-peak amplitude

To determine peaks, the global minimum in the interval [29 ms 114 ms] was assigned to be the observable ‘N75 latency’; the global maximum in the interval [86 ms 143 ms] was assigned to be ‘P100 latency’ (marked by icons in Fig. 1). The difference between the

![Fig. 1. Two VEP waveforms recorded from each subject. First session (black line, circles), second session after 8 months (gray line, squares), positive up. Subjects are sorted by decreasing VEP amplitude of the first session. The first icon (circle or square) indicates N75 latency and the second icon indicates P100 latency. The difference in amplitude between the two icons of each waveform is defined as the peak-to-peak amplitude of the VEP waveform. The similarity of two waveforms recorded in two sessions of the same subject is characterized by the t-value tshape.](image-url)
values of the VEP waveform at the two latencies was taken as peak-to-peak amplitude.

2.4. Similarity in shape of VEP waveforms

To assess the similarity in shape of VEP waveforms, we performed pairwise comparisons between the waveforms \( k \) and \( m \), where indices \( k, m \) run over all VEP recording sessions. The waveforms are written as \( x^{(k)}_i \) and \( x^{(m)}_i \) where \( i \) denotes one of \( n \) time points. The waveforms were regressed pairwise onto each other using the model

\[
x^{(k)}_i = \beta_1 + \beta_2 \cdot x^{(m)}_i + e_i, \quad i = 1, \ldots, n
\]

where \( e_i \) denotes random noise. We included the whole time range \([0...292 \text{ ms}] \) in the analysis (Fig. 1). The ordinary least-squares estimation was used in the linear regression model. If the shape of two waveforms is similar, the slope \( \beta_2 \) is nearly to one and the standard error \( SE \) of coefficient \( \beta_2 \) is near to zero. The \( t \)-value

\[
t_{\text{shape}} = \frac{\beta_2}{SE}
\]

describes the strength of the relationship between the two waveforms. In this way, we generate the observable \( t_{\text{shape}} \) describing the similarity in shape of two VEP waveforms. The observable \( t_{\text{shape}} \) is unaffected by a shift in VEP amplitude as it may occur, e.g., due to uncertainties in electrode placement.

While we have chosen here a regression model with \( t_{\text{shape}} \), the use of linear correlation to quantify similarity of waveforms leads to comparable results for subject recognition.

2.5. Generalized linear regression model

To describe the similarity between two VEP recording sessions, four observables were derived from the VEP waveforms. The differences in Z-scores of N75 latency (\( \Delta \text{N75} \)), P100 latency (\( \Delta \text{P100} \)), and peak-to-peak amplitude (\( \Delta \text{APP} \)) describe the similarity of the peak values. The \( t \)-value of the linear regression (\( t_{\text{shape}} \)) assesses the similarity in shape of the VEP waveforms. The four observables were entered in a generalized linear regression model (GLM).

Since every observation describes a comparison of two VEP recording sessions, the response variable denotes whether it is an intra- or inter-individual comparison and is therefore binomially distributed. This leads to a GLM with “Logit” as canonical link function for binary response variables:

\[
y_i = \frac{\exp(x_i \mathbf{\beta})}{1 + \exp(x_i \mathbf{\beta})} + E_i
\]

\[
\mathbf{\beta} \cdot X = \beta_0 + \sum_{j=1}^{4} \beta_j \cdot x^{(j)}_i
\]

\( \mathbf{\beta} \cdot X \), linear combination of coefficient vector \( \mathbf{\beta} \) with predictors \( X \); \( x^{(j)}_i \), the 4 elements of predictors \( X \) are \( t_{\text{shape}}, \Delta \text{N75}, \Delta \text{P100} \) and \( \Delta \text{APP} \); \( y_i \), response variable for all pairwise comparisons \( i \). It describes whether the comparison is intra- or inter-individual (\( y_i = 1 \)), if the two recording sessions are from the same subject, \( y_i = 0 \) otherwise; \( E_i \), random errors.

We compared the first and second session pairwise with all sessions of all subjects. The comparison with the highest recognition probability \( y_i \) indicates recognition of a subject. Over all 20 sessions, we obtain an estimated sensitivity for recognition. The estimated sensitivity is taken as probability of success in a binomial trial based on the number of correctly assigned sessions observed in 20 sessions. Throughout the text we present the sensitivity estimates and the 95% confidence intervals.

To test its stability and to avoid over-fitting, the GLM was fitted for each subject separately under exclusion of the data of this subject (cross-validation by the leave-one-out method). This means that comparisons including subject \( k \) were not used to estimate coefficients, which predict the recognition probability for subject \( k \).

3. Results

Fig. 1 shows individual VEP waveforms of the 10 subjects. Subjects were sorted in order of descending VEP amplitude. Waveform differences between subjects (inter-individual variability) and between test session (solid lines) and retest session (gray lines) can be detected by visual inspection.

In a cross-sectional approach, Fig. 2 shows VEP latencies and VEP amplitudes for both sessions, at baseline and 8 mo, of all subjects. The distributions of N75 latency (mean 71 ± 10 ms) and P100 latency (mean 106 ± 5 ms) are significantly different from the normal distribution \( (p < 0.01, \text{Lilliefors test}) \), in line with the VEP standard (Odom et al., 2004). The peak-to-peak amplitude \( (3.1 ± 1.2 \, \mu V) \) is not significantly different from the normal distribution. None of the observables showed a correlation with the age of our subjects.

In a longitudinal approach, we first calculated the linear correlation \( \rho \) and the repeatability coefficient \( rc \) (Bland and Altman,
Fig. 3. Bland–Altman plot of latencies N75 and P100 and amplitudes. For N75 latency, P100 latency and VEP peak-to-peak amplitude, averages of first and second recording sessions are plotted against differences between first and second recording sessions for each subject. The coefficient of repeatability (rc) equals 2 SD and is indicated by dotted lines.

Fig. 4. Similarity of VEP waveform. Similarity of VEP traces of subject 3 (sessions 3 and 13). The traces of the first session (3) and second session (13) are regressed onto each other. Each marker represents a time point. The two traces are correlated with $\rho = 0.97$. The slope $b$ of the regression line (dashed) and the standard error SE give the value $t_{\text{shape}} = b/SE = 129$.

Fig. 5. Subject recognition. Matrix of $t$-values for comparisons between all VEP recording sessions. (a) Values of $t_{\text{shape}}$ are highest in the main diagonal which represents the trivial case. (b) For better visibility, only maximal values of $t_{\text{shape}}$ of each row are plotted after exclusion of the main diagonal. In the first row (subject 1), the position of the maximum at column 11 indicates that the second session of this subject was correctly assigned. Maximal values of $t_{\text{shape}}$ in the diagonals 10 above/below the main diagonal indicate correct recognition. Row 6 with maximal value of $t_{\text{shape}}$ off these diagonals represents the single session, which was not correctly assigned.
Fig. 6. Stability of the GLM. To cross-validate the recognition probabilities, one GLM is fitted for each session under exclusion of the data of session $k$. A value of $t_{\text{GLM}}$ exceeding $\pm 1.96$ corresponds to an error probability of $p < 0.05$. The two curves show the value of $t_{\text{GLM}}$ for parameters $\beta_0$, which reflects the constant term, and $\beta_1$, which reflects the similarity of shape $t_{\text{shape}}$.

ima in the main diagonal represent the trivial case of self-recognition of each session. Fig. 5b shows the maxima of each row after exclusion of the main diagonal. Maximal values of $t_{\text{shape}}$ in the diagonals 10 above and below the main diagonal indicate correct recognition. The session in row 6 is not correctly assigned. The number of sessions correctly assigned as in Fig. 5b. The high recognition rate achieved with $t_{\text{shape}}$ indicates that the similarity of VEP waveform is high compared to its inter-individual variability. In other approaches to VEP variability, short-term test–retest reliability of VEP observables was used to discriminate between a control group and groups of patients with ophthalmological disorders (Parisi et al., 2006) or the VEP variability was analyzed in terms of principal components (Dandekar et al., 2007). In view of the large variability of VEP amplitudes we propose the $t_{\text{shape}}$ method to analyze the general waveform of VEP for longitudinal studies.

4. Discussion

As indicated by the high number of correctly recognized subjects, the overall waveform similarity $t_{\text{shape}}$ is the most relevant observable for test–retest reliability of the VEP. The other observables are ineffective predictors for recognition even though they are highly correlated between recording sessions (Table 1). This finding supports the general concept that correlation is not applicable to determine repeatability of measurements (Bland and Altman, 1986). Correlation values may, however, serve to compare our results to those in the literature (Table 1). The correlation values found here are possibly higher than those in earlier reports (Hall et al., 1973; Schellberg et al., 1987). The P100 latency repeatability coefficient of 9.0 ms is in a similar range as the reported P100 variability of 11 ms (Oken et al., 1987). In a study with intersession-interval of up to 28 days, the high variability of VEP amplitudes was attributed to the mood of subjects (Joost et al., 1992). The statistical approach presented here has earlier been applied to spectral EEG observables, where a similar sensitivity for subject recognition was obtained (Näpflin et al., 2007, 2008).

The high recognition rate achieved with $t_{\text{shape}}$ indicates that intra-individual stability of VEP waveform is high compared to its inter-individual variability. In other approaches to VEP variability, short-term test–retest reliability of VEP observables was used to discriminate between a control group and groups of patients with ophthalmological disorders (Parisi et al., 2006) or the VEP variability was analyzed in terms of principal components (Dandekar et al., 2007). In view of the large variability of VEP amplitudes we propose the $t_{\text{shape}}$ method to analyze the general waveform of VEP for longitudinal studies.

4.1. Efficiency of the statistical model

The GLM provides a method to cross-validate the sensitivity of recognition. The high recognition rate indicates that the similarity of VEP shape can be encoded by $t_{\text{shape}}$ in a GLM with high statistical efficiency. Since $t_{\text{GLM}}$ values (Fig. 6) were obtained by leaving one session $k$ ($k = 1 \ldots 20$) out of the estimation process, the stability of the obtained $t_{\text{GLM}}$ documents that the model is resistant against the exclusion of sessions and thereby that the group of sessions (number of observations for the GLM) is sufficiently large. In general, redundant information leads to over-fitting in our true out-of-sample design, thus constraining us to focus on a small number of observables that are statistically significant for our data. This becomes apparent in the optimal sensitivity of the method for inclu-

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Comparison of measures of test–retest reliability.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured quantity</td>
<td>Number of subjects</td>
</tr>
<tr>
<td>Present study</td>
<td>VEP</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Hall et al. (1973)</td>
<td>VEP</td>
</tr>
<tr>
<td>Oken et al. (1987)</td>
<td>VEP</td>
</tr>
<tr>
<td>Schellberg et al. (1987)</td>
<td>VEP</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Gasser et al. (1985)</td>
<td>Resting EEG spectra</td>
</tr>
<tr>
<td>Näpflin et al. (2007)</td>
<td>Resting EEG spectra</td>
</tr>
<tr>
<td>Näpflin et al. (2008)</td>
<td>Task EEG spectra</td>
</tr>
</tbody>
</table>
sion of observable $t_{\text{shape}}$ in the GLM. Since the inclusion of the other observables $\Delta N75$, $\Delta P100$ and $\Delta P300$ reduces sensitivity and/or $t_{\text{GLM}}$, these observables seem not to be relevant to the question of test–retest reliability. As the results of the statistical model were cross-validated with the leave-one-out method, the results obtained are not only valid for the sample of subjects in this study but also for the basic population and thereby achieve high explanatory power.

4.2. What is signal, what is noise?

Our method demonstrates high test–retest reliability of the VEP waveforms: the intra-individual variability of the waveform is low compared to its inter-individual variability. As can be seen from Fig. 1, the variability across subjects is indeed high and in some cases very different from standard VEPs (Odom et al., 2004). The large variability raises the question on what is ‘true VEP signal’ as opposed to noise obscuring the true VEP signal.

We have sorted subjects by VEP amplitude so that we can comment on the signal to noise relationship. The VEP amplitude in subjects 6...10 is below the median (Fig. 2c). In this low-amplitude subject group, inter-individual variability in latency estimates is relatively high (Fig. 2a and b). Also the similarity in waveform $t_{\text{shape}}$ between first and second session of the same subject is lower for the low-amplitude subject group than for the high-amplitude subjects 1...5 (Fig. 2d, $p = 0.02$, Wilcoxon ranksum test). At first glance, one might therefore attribute a larger level of ‘noise’ to waveforms with lower VEP amplitude. However, intra-individual variability in latencies and amplitudes is not necessarily higher for all of the low-amplitude subjects (Fig. 2). Furthermore, the intra-individual values of $t_{\text{shape}}$ are still large in the framework of test–retest reliability as can be seen in Fig. 1 and Fig. 5b. Indeed, nine of the ten sessions of the low-amplitude subject group were correctly assigned. This indicates that also VEP waveforms with low-amplitude contain sufficient information in their waveform for successful recognition of individual subjects.

5. Conclusions

The intra-individual variability of VEP waveform in women of childbearing age was low compared to the inter-individual variability, as measured over a test–retest interval of 8 months. Contrary to earlier reports in the literature, the recognition rate was high with our statistical approach and comparable to that of resting EEG with eyes closed. Cross-validation allows to single out VEP waveform as the most relevant observable for recognition. The high intra-individual stability of VEP waveform of repeated measurements in the same subject makes it possible to study the influence of external factors longitudinally in individuals. The long interval of high intra-individual stability provides a statistical baseline against which effects on the VEP can be contrasted, may they be effects of progressing disease, or effects of clinical therapy.

Acknowledgements

We thank the participating women, Jasmin Tajeri Foman for technical assistance and Matthias Sitzler for helping to unravel the binary VEP data format. The study was funded by Unilever Food and Health Research Institute, Vlaardingen, Netherlands and the Swiss Federal Institute of Technology Zurich, Switzerland.

References