Test–retest reliability of P50, N100 and P200 auditory sensory gating in healthy subjects

Johannes Rentzsch a,b,⁎, Maria C. Jockers-Scherübl a,b, Nash N. Boutros d, Jürgen Gallinat a,c

a Department of Psychiatry, Charité — Universitätsmedizin Berlin, Germany
b Campus Benjamin Franklin, Germany
c Campus Mitte, Germany
d Wayne State University School of Medicine, Departments of Psychiatry and Behavioral Neurosciences, Detroit, Michigan, USA

Received 18 February 2007; received in revised form 28 July 2007; accepted 18 October 2007
Available online 25 October 2007

Abstract

Rationale: Suppression of middle latency auditory evoked responses is considered an index for the multistage sensory gating process. This has been observed in sequentially occurring P50, N100 and P200 components in a dual-click procedure. Since P50 sensory gating deficits have been observed in schizophrenic patients and first degree relatives, this parameter was suggested as an intermediate phenotype of the disease. However, most studies only show a low reliability for P50 sensory gating and neither N100 nor P200 sensory gating have been sufficiently tested.

Methods: Reliability of P50, N100 and P200 sensory gating was measured in 41 healthy subjects in two sessions, four weeks apart, using intra-class correlation. Sensory gating was calculated as ratio-gating (second response magnitude/first response magnitude × 100) as well as difference-gating (first response magnitude minus second response magnitude).

Results: The difference-gating showed good to excellent reliabilities independently of the amplitude-measurement method applied (P50 peak-to-peak 0.75 and baseline-to-peak 0.74, N100 peak-to-peak 0.63 and baseline-to-peak 0.70, P200 peak-to-peak 0.82 and baseline-to-peak 0.79). Regarding ratio-gating, best temporal stability was observed for the P200 (peak-to-peak 0.58 and baseline-to-peak 0.62). Reliability of P50 ratio-gating strongly depends on the amplitude-measurement method (peak-to-peak 0.0 and baseline-to-peak 0.46).

Conclusion: Regarding long-term reliability in healthy subjects the difference-gating of all three evoked responses and the ratio-gating of the P200 component may be useful tools for clinical or intermediate phenotype studies measuring different stages of the auditory sensory gating process. In contrast, the reliability of the P50 and N100 ratio-gating component seems to be insufficient for this purpose. However, long-term reliability remains to be confirmed in clinical samples.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Sensory gating; Auditory evoked potential; Reliability; Long-term; P50; N100; P200

1. Introduction

Deficits in inhibitory functions under various experimental conditions are often found in schizophrenic patients and may be considered a core feature of the schizophrenic pathophysiology (Boutros et al., 2004; Braff et al., 2001; Clementz, 1998; Curtis et al., 2001; Mathalon et al., 2002; Moritz et al., 2001; Poole et al., 1999). A disturbed inhibitory function of the brain to irrelevant sensory input, i.e. an impaired ability to filter out irrelevant sensory information, may be related to an inadequate sensory information processing and result in a flooding of higher cortical centers with irrelevant information (Boutros et al., 2004). Such impaired sensory filtering, shown by failure to suppress the evoked potentials elicited by the second of an identical stimulus pair, has been observed in auditory paired stimulus paradigms in schizophrenia and is called ‘sensory gating’. It is measured as a ratio of the second to the first evoked
response amplitude. While a low ratio indicates an effective gating or filtering out of irrelevant sensory information, schizophrenic patients often show noticeably higher ratios than healthy controls; see Boutros and Aysenil (1999) for discussion of the theories related to the sensory gating phenomenon. Sensory gating deficits in schizophrenia were mostly investigated for P50 response, i.e. those evoked 50 ms after stimulus onset (Adler et al., 1982; Baker et al., 1990; Boutros et al., 1993; Freedman et al., 1991; Griffith et al., 1998; Hsieh et al., 2004; Rentzsch et al., 2007; Yee et al., 1998).

The deficit in P50 sensory gating ratio has been considered an intermediate phenotype of schizophrenia because it (1) is associated with the disease (Bramon et al., 2004); (2) is largely unaffected by its course and psychopathology (Adler et al., 1990a; Baker et al., 1987; Boutros et al., 1999; Boutros et al., 2004; Ward et al., 1996; Yee et al., 1998), even though some studies found normal P50 ratio scores in paranoid compared to undifferentiated schizophrenic patients (Boutros et al., 1991a); (3) has been shown to be heritable (Myles-Worsley et al., 1996; Young et al., 1996) and co-segregated with schizophrenia within families (Clementz et al., 1998b; Siegel et al., 1984; Waldo et al., 1991); and (4) is found in non-schizophrenic family members at a higher rate than in the general population (Clementz et al., 1998b; Myles-Worsley, 2002; Siegel et al., 1984; Waldo et al., 1995).

Apart from the criteria listed above, a further requirement for defining intermediate phenotypes of schizophrenia is sufficient test–retest reliability (Bearden and Freimer, 2006; De Geus and Boomsma, 2001; Gooding et al., 2004; Gottesman and Gould, 2003). However, only a low temporal stability of the P50 sensory gating measured as ratio scores could be shown in most of the retest investigations. In healthy subjects, reports of within-session retest stability and between-session retest stability of up to two weeks have ranged from 0 to 0.3 (Boutros et al., 1991a,b; Clementz et al., 1997; Kathmann and Engel, 1990; Naber et al., 1992; Smith et al., 1994). Better test–retest correlations (0.47 and 0.66) were reported by Lamberti et al. (1993) and in a more recent study by Hall et al. (2006). However, in a re-examination of Lamberti’s data by Smith et al. (1994) the correlations ranged from 0.1 to 0.5. The P50 sensory gating retest reliability in schizophrenic patients has barely been studied. For instance, Kathmann and Engel (1990) found a very poor retest stability, with a negative Pearson correlation of –0.2 in schizophrenic patients and Clementz et al. (1998a) found a similar retest reliability for healthy controls and schizophrenic patients, with the negative Pearson correlation amounting to 0.09. On the whole, these studies suggest that P50 sensory gating only shows low to, at best, moderate retest reliability. This questions the usefulness of the P50 sensory gating as an intermediate phenotype when measured as a ratio of second to first P50 amplitude.

Auditory sensory gating is a multistage operation and not only occurs in P50 response, but also in the subsequent N100 and P200 responses. These later components have been related to higher-order cognitive processes (Kisley and Comwell, 2006; Gallinat et al., 2002). Boutros et al. (2004) suggested that P50, N100 and P200 responses reflected different stages of information processing (pre-attentive: reflected by the P50; early attentive: reflected by the N100; later attentive: reflected by the P200). At the levels of N100 and P200 responses dysfunctional gating has also been observed in schizophrenic patients (Boutros et al., 1999b; Clementz and Blumenfeld, 2001; Nagamoto et al., 1989). A recent twin study observed a strong heritability of these components (Anokhin et al., 2006). Thus, there is evidence supporting N100 and P200 sensory gating also as intermediate phenotypes for schizophrenia. However, their retest reliability has not yet been adequately evaluated (Fuerst et al., 2007; Smith et al., 1994).

The aim of the study was to investigate the retest stability of P50, N100 and P200 sensory gating in healthy subjects. Unlike most investigations, a relatively long interval of four weeks was imposed between both recordings and a large sample of 41 carefully screened healthy subjects was investigated. We also measured difference-gating to identify the most reliable parameter (Hall et al., 2006; Jerger et al., 1992; Smith et al., 1994).

In contrast to most P50 sensory gating studies we applied much shorter inter-pair intervals, because there is evidence suggesting that a marked P50 sensory gating can also occur at much shorter inter-pair intervals. Dolu et al. (2001) investigated P50 sensory gating with the paired stimulus design using the long inter-pair interval of 8 s while applying different inter-stimulus intervals. As expected they found a marked gating phenomenon when the inter-stimulus interval was 0.5 s (P50 sensory gating ratios 37%), but no gating phenomenon was found at inter-stimulus intervals of 0.75 s and 1 s (P50 sensory gating ratios 114% and 93%, respectively). This implies that a full recovery of the P50 potential is possible even after 1 or 2 s.

2. Material and methods

2.1. Subjects

The study was approved by the ethics committee of the Charité - Universitätsmedizin Berlin, Campus Benjamin Franklin. All subjects gave written, informed consent. Subjects were recruited by newspaper advertisements and were remunerated for their participation. First, the subject’s health status was screened during a telephone interview carried out by a trained medical student using a structured questionnaire and then further examined by an experienced psychiatrist with a structured interview (M.I.N.I.) (Sheehan et al., 1998). Exclusion criteria were any Axis I diagnoses following DSM-IV. Further, severe internal or neurological diseases (such as migraine, Parkinson, ischemic brain insults, non-compensated hypothyroidism or diabetes mellitus), hearing disorders or consumption of psychotropic medication. Any of these led to exclusion. None of the subjects reported a family history of schizophrenia or schizoaffective disease. 18 males and 23 females (36 right-handed, 2 left-handed, 3 amidedextrous; years of overall education: mean±SD: 14.9±2.0, range: 10–18) aged between 19 and 51 years (mean±SD: 34.0±8.9) were included — all of whom were tested twice in a mean interval of 28±2.2 days (range 23–33).

2.2. Task and ERP-recording

Subjects were seated in a sound-attenuated and electrically shielded room adjacent to the recording apparatus (Neuroscan
Table 1
Latencies (ms), amplitudes (μV) in response to 1st and 2nd stimulus and gating indices for P50, N100 and P200 for test session 1 (t1) and 2 (t2) for both amplitude measurements (bp: baseline-to-peak amplitude; pp: peak-to-peak amplitude [N40–P50, N40–N100, N40–P200, respectively])

<table>
<thead>
<tr>
<th>Peak indicesa</th>
<th>P50 mean (SD)</th>
<th>N100 mean (SD)</th>
<th>P200 mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency</td>
<td>t1</td>
<td>t2</td>
<td>t1</td>
</tr>
<tr>
<td>1st</td>
<td>59.9 (2.9)</td>
<td>60.4 (3.0)</td>
<td>102.7 (15.5)</td>
</tr>
<tr>
<td>2nd</td>
<td>57.7 (3.1)</td>
<td>57.7 (3.5)</td>
<td>83.8 (13.6)</td>
</tr>
<tr>
<td>bp amplitude</td>
<td>t1</td>
<td>t2</td>
<td>t1</td>
</tr>
<tr>
<td>1st</td>
<td>3.0 (0.7)</td>
<td>2.8 (1.6)</td>
<td>−5.2 (2.5)*</td>
</tr>
<tr>
<td>2nd</td>
<td>0.9 (0.69)</td>
<td>0.9 (0.6)</td>
<td>−2.5 (1.2)*</td>
</tr>
<tr>
<td>pp amplitude</td>
<td>t1</td>
<td>t2</td>
<td>t1</td>
</tr>
<tr>
<td>1st</td>
<td>3.3 (2.1)</td>
<td>3.1 (1.9)</td>
<td>5.1 (2.7)</td>
</tr>
<tr>
<td>2nd</td>
<td>1.1 (0.8)</td>
<td>1.1 (0.8)</td>
<td>1.9 (1.3)</td>
</tr>
</tbody>
</table>

Gating indicesa

| bp ratio       | 34.9 (19.8)   | 37.6 (25.9)   | 53.4 (28.9)   | 47.9 (39.5)   | 44.5 (31.4)   | 48.1 (34.5)   |
| pp ratio       | 36.6 (21.4)   | 35.8 (23.2)   | 49.2 (39.6)   | 54.1 (54.9)   | 57.3 (43.4)   | 53.7 (40.1)   |
| bp difference  | 2.1 (1.3)     | 1.9 (1.3)     | 2.7 (2.1)     | 2.7 (2.1)     | 3.4 (2.6)     | 3.3 (2.8)     |
| pp difference  | 2.1 (1.4)     | 2.1 (1.4)     | 3.1 (2.3)     | 2.7 (2.3)     | 3.1 (2.91)    | 3.3 (3.0)     |

Correlation between baseline-to-peak and peak-to-peak measurements for t1: P50: 1st 0.95**, 2nd 0.86**, ratio-gating: 0.53**, difference-gating: 0.93**; N100: 1st 0.94**, 2nd 0.36*, ratio-gating: 0.4*, difference-gating: 0.79**; P200: 1st 0.94**, 2nd 0.61*, ratio-gating: 0.58*, difference-gating: 0.88**.

⁎⁎ p<0.001, *p<0.05, all other p>0.1.

a Paired t-test test session 1 (t1) and 2 (t2) without correction for multiple testing.

SynAmps model 5083, El Paso, TX), with closed eyes, in a slightly reclined chair with a headrest. Auditory stimuli consisted of 175 identical pairs of clicks generated by a PC-stimulator with “Creative Labs SoundBlaster 16” soundcard (duration: 1 ms square wave, 109 dB1) through calibrated headphones. Paired clicks were separated by 500 ms. Four different fixed inter-pair intervals were used in a pseudo-randomized order (1.5 s, 3 s, 3.8 s, and 4.6 s; mean inter-pair interval: 2.8 s).

The responses evoked by clicks were recorded with 29 tin electrodes referred to Cz, using an electrode cap following the International 10/20 System with additional electrodes. Fpz served as ground. Eye movements were recorded across an electrode placed 1 cm laterally to the eye (EOG). Electrode impedance was less than 10 kΩ. Data were collected at a sampling rate of 500 Hz (gain 5000; analog band pass filter: 0.15–100 Hz). The EEG measures were taken by the same investigator on both occasions.

2.3 Parameterization and peak definition

The EEG was analyzed offline using “Brain vision analyzer” software (Version 1.1, Brain vision, Munich, Germany). The data were digitally filtered (P50: high-pass 10 Hz, 24 dB octave; N100 and P200: high-pass 0.5 Hz, 24 dB octave) after re-referencing to average reference, segmentation (350 ms pre-first-stimulus to 800 ms post-second-stimulus), artefact rejection (i.e. exclusion of segments showing activity greater/ lower than 100 μV in any of the 29 channels and/or in the EOG at any point of the sweep period; no further blink correction algorithm was applied) and baseline-correction. After averaging of the remaining sweeps, latencies and amplitudes of the P30, N40, P50, N100 and P200 at the Cz electrode were analyzed on the basis of automatic peak detection in combination with a visual control blind to the time point of testing. All subjects had more than the minimum of 90 segments per average. The P50 component was defined as the most positive response between 40 and 80 ms post-stimulus preceded by a P30 wave in a 20–50 ms range. If there was no identifiable P30, the most prominent positive component in the P50 time range was used as P50. The N100 and P200 component was defined as a prominent negative–positive complex (N100: 60–170 ms, P200: 100–260 ms). All amplitudes were measured both in relation to the pre-stimulus baseline (baseline-to-peak; referred to as P50, N100 and P200) and in relation to the N40 peak, which served as a baseline (peak-to-peak; referred to as N40–P50, N40–N100 and N40–P200). N40 was defined as the most prominent negative peak between P30 and P50. Whenever these were equivocal, the most negative peak preceding the P50 was used as N40.

When no amplitude was identifiable for the first stimulus, the subject’s response was excluded from further analysis (P50: one subject; N100: no subject; P200: one subject). If this was the case for the second stimulus, it was interpreted as maximum suppression and the amplitude was set to zero in accordance.

1 Loudness calibration was performed by electrical measurements of continuous sine-wave voltages at the headphone’s terminals using calibrated headphones. The given sound level in dB of the clicks corresponded to the sound pressure level in dB of a sine wave with equivalent peak amplitude. Most auditory sensory gating studies were done using lower click sound levels, but higher values have also been used, e.g. 120 dB by Ringel et al. (2004). However, in most studies no technical information on the measurement method of the ‘click sound level’ was given. Sound level values for clicks cannot be compared without defining the measurement procedure precisely.

2 As the peak-to-peak amplitude measurement is common in P50 literature (called trough-to-peak), the N100 and P200 amplitudes are usually measured in relation to the preceding peak, which is the P50 for the N100 and the N100 for the P200. As both the P50 and the N100 responses show a gating phenomenon, this interferes with the gating estimation of the following responses. Hence the N40, which shows no gating phenomenon (Arnfred et al., 2001) is a better reference for N100 and P200 amplitude measurements. Furthermore, this results in a common reference for all three responses.
with Dolu et al. (2001) and Nagamoto et al. (1989) (this was the case in five P50 responses; three N100 responses; five P200 responses). Gating indices were calculated as ratio-gating (2nd stimulus amplitude / 1st stimulus amplitude \times 100) as well as difference-gating (1st stimulus amplitude − 2nd stimulus amplitude) for both baseline-to-peak and peak-to-peak amplitudes.

2.4. Statistics

All analyses were conducted using “Statistical Package for Social Sciences” (SPSS version 13.0 for Windows; SPSS Inc., Chicago, IL). All parameters are presented as mean and standard deviation. To estimate the normal distribution we used the one-sample Kolmogorov–Smirnov test. Except for the N100-2nd latency of test session 1 and for the ratio-gating of the N100 of test session 2, the normal distribution was a good fit for all other evoked parameters (\( p > 0.05 \)). Evoked responses and gating indices between test session 1 and test session 2 and between 1st and 2nd stimulus were compared using paired-samples t-tests. To determine test–retest reliability, intra-class correlation coefficients (ICC) were used in a two-way mixed effect model (single measure) with consistence: Model (3.1). ICCs were used because they provide a better measurement of true agreement than simple Pearson’s correlations (for discussion see e.g. Farahat et al., 2003). The ICC depends on inter-subject and intra-subject variability. The ICC will be high when the intra-subject variability is low relative to the total of inter-subject and intra-subject variability. The ICC will be low when intra-subject variability is high relative to this total variability. For example, an ICC of 0.60 for test and retest measurements means that 60% of the variability is caused by inter-subject variability, whereas 40% is caused by intra-subject variability (Rouse et al., 2004). Standards for assessing the degree of test–retest agreement were adopted from previous practice (Fendrich et al., 1990; Patten et al., 2003; Rouse et al., 2004; Stokdijk et al., 2000) in accordance with Fleiss (1986). That is, a reliability coefficient of less than 0.39 would be considered poor, 0.4–0.59 fair, 0.6–0.75 good, and greater than 0.75 excellent. However, it has to pointed out, that classification of the level of reliability, represented by the ICC index, is often an area of disagreement among different researchers, e.g.: upper level of ‘poor’ ranged from 0.2 to 0.69, ‘fair’ from 0.4 to 0.79 and ‘good’ from 0.6 to 0.8; for overview see Table 1 and page 8 in Hager, 2003 [Master thesis]). Other adjectives to describe ranges of reliability values were provided e.g. by Shrout (1998), but with mostly similarly ranges as we used (0.0–0.1 virtually none, 0.11–0.40 slight, 0.41–0.60 fair, 0.61–
latencies and the N100 amplitudes were observed (see Table 1). Meters at test session 1 and test session 2, except for the P200 the second stimulus compared to the first one (for all responses 1st) to the second stimulus (P50-2nd, N100-2nd, P200-2nd) (for all responses in both test sessions. As indicated by ratio and difference values, all responses show a marked and significant reduction in the amplitude from the first (P50-1st, N100-1st, P200-1st) to the second stimulus compared to the first one (for all responses p < 0.0001). Fig. 2 depicts a boxplot for ratio-gating of the three responses. No significant differences between the evoked parameters at test session 1 and test session 2, except for the P200 latencies and the N100 amplitudes were observed (see Table 1).

The Pearson correlations of P50, N100 and P200 sensory gating with one another (test session 1) were not significant for ratio-gating (baseline-to-peak: rP50–N100 = 0.22, rP50–P200 = 0.20, rN100–P200 = 0.08; peak-to-peak: rP50–N100 = 0.02, rP50–P200 = 0.02, rN100–P200 = 0.21; for all p ≥ 0.1) but significant for most difference-gating values (baseline-to-peak: rP50–N100 = 0.39, rP50–P200 = 0.36, rN100–P200 = 0.36 peak-to-peak: rP50–N100 = 0.37, rP50–P200 = 0.33; all p ≤ 0.05, but not significant for rN100–P200 = 0.12).

3.2. Test–retest reliabilities of latencies, amplitudes and gating

Ratio-gating as well as difference-gating data of test sessions 1 and 2 for all three responses are given as scatterplot in Fig. 3. The intra-class correlation coefficients (ICC) as an index for reliability between both test sessions are shown in Table 2. For seven of the 30 tested parameters the ICCs were above 0.75 and for 13 parameters the ICCs were within 0.6 and 0.75. This was the case for all, but one amplitude (N100-2nd) and for all difference-gating values. For latency and ratio-gating this was only the case in two of six (P50-1st, N100-2nd) and in one of six (P200 ratio-gating) measurements (see Table 2). ICCs of amplitudes showed no marked differences between the two methods of amplitude measurement (pp, bp) except for the second N100 amplitude (ICCpp: 0.34 and ICCbp: 0.62). Since the N40 served as baseline for peak-to-peak amplitude measurements its quality affects not only the quality of the peak-to-peak amplitudes but also of the peak-to-peak gating. Thus, we calculated the reliability for N40 of the P50 and of the later N100 and P200 responses. The ICCs for the first N40 were higher than for the second one: N40 baseline of the N40–P50 amplitude: N40-1st: 0.84 (F = 11.72; 95% CI: 0.72–0.91; p < 0.0001), N40-2nd: 0.52 (F = 3.14; 95% CI: 0.25–0.71; p < 0.0001) and N40 baseline of the N40–N100 and N40–P200 amplitude: N40-1st: 0.68 (F = 5.25; 95% CI: 0.47–0.86; p < 0.0001), N40-2nd: 0.58 (F = 3.72; 95% CI: 0.33–0.75; p < 0.0001).

For ratio-gating the P200 showed the highest ICCs (0.62 and 0.58), followed by the ICCs of P50 (ICC 0.47) and N100 (0.23 and 0.35) with the lowest ICC observed for the N40–P50 (ICC = 0.03). Note, that the amplitude-measurement method had a marked impact on the ratio-gating ICC of the P50 component but not on ratio-gating ICC of the N100 and the P200 components. The ICCs for difference-gating were comparable between the three responses and amplitude-measurement methods, highest again for P200 (0.79 and 0.82), followed by P50 (0.74 and 0.75) and N100 component (0.7 and 0.63) (see Table 2).

Concerning the low ICC of the N40–P50 ratio-gating we performed further analyses. Recommend by Bland and Altman (1986), we plotted the difference test session 1 minus test session 2 of N40–P50 ratio-gating against the mean of the N40–P50 ratio-gating of both test sessions (Bland–Altman plot, not shown). This plot allows to detect outliers, and also indicates...
whether variation of N40–P50 ratio-gating increases or decreases as a function of N40–P50 ratio-gating. The plot shows that the difference between N40–P50 ratio-gating test session 2 and test session 1 was near zero with a roughly constant variation across the range of values (mean±SD: 0.79±32.2; limits of agreement at −63.5 and 65.1). There was no significant correlation between the difference of ratio-gating test session 1 and test session 2, and the mean of ratio-gating test session 1 and 2 (r = −0.08; p = 0.62). Thus, variations of N40–P50 ratio-gating across both test sessions were not a function of the ratio-values. Five outliers were identified whose differences were not within two standard deviations of the mean difference. Exclusion of these subjects results in a slightly better ICC of N40–P50 ratio-gating (ICC: 0.39; F = 2.29; 95% CI: 0.07−0.64; p < 0.01). These subjects showed no significant difference in age, sex and number of P50 sweeps. In four out of these five outliers the amplitudes for test session 1 and test session 2 were smaller than 1 μV. When all 40 cases were considered, seven individuals for test session 1 and nine individuals for test session 2 had a N40–P50-1st amplitude smaller than 1 μV. When these subjects were excluded the ICC for the remaining 31 subjects increased slightly (ICC: 0.46; F = 2.68; 95% CI: 0.13−0.7; p = 0.004).

Table 2
Intra-class correlation test sessions 1 and 2 (ICCt1–t2): latencies, amplitudes to 1st and 2nd stimulus and gating indices for both amplitude measurements (bp: baseline-to-peak amplitude; pp: peak-to-peak amplitude [N40–P50, N40–N100, N40–P200, respectively])

<table>
<thead>
<tr>
<th>Peak indices</th>
<th>P50</th>
<th>N100</th>
<th>P200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICCt1–t2 (F; CI)</td>
<td>ICCt1–t2 (F; CI)</td>
<td>ICCt1–t2 (F; CI)</td>
</tr>
<tr>
<td>Latency 1st</td>
<td>0.73 (6.53; 0.55–0.85)**</td>
<td>0.54 (3.33; 0.28–0.72)**</td>
<td>0.55 (3.48; 0.3–0.74)**</td>
</tr>
<tr>
<td>Latency 2nd</td>
<td>0.58 (3.80; 0.34–0.76)**</td>
<td>0.71 (5.69; 0.54–0.83)**</td>
<td>0.51 (3.09; 0.24–0.71)**</td>
</tr>
<tr>
<td>bp amplitude 1st</td>
<td>0.86 (13.10; 0.75–0.92)**</td>
<td>0.71 (7.39; 0.59–0.87)**</td>
<td>0.82 (9.97; 0.68–0.90)**</td>
</tr>
<tr>
<td>bp amplitude 2nd</td>
<td>0.73 (6.33; 0.54–0.85)**</td>
<td>0.34 (2.17; 0.07–0.61)**</td>
<td>0.70 (5.72; 0.50–0.83)**</td>
</tr>
<tr>
<td>pp amplitude 1st</td>
<td>0.89 (17.29; 0.80–0.94)**</td>
<td>0.70 (5.64; 0.50–0.83)**</td>
<td>0.78 (8.24; 0.63–0.88)**</td>
</tr>
<tr>
<td>pp amplitude 2nd</td>
<td>0.78 (7.93; 0.62–0.88)**</td>
<td>0.62 (4.27; 0.39–0.78)**</td>
<td>0.73 (6.48; 0.55–0.85)**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gating indices</th>
<th>P50</th>
<th>N100</th>
<th>P200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ICCt1–t2 (F; CI)</td>
<td>ICCt1–t2 (F; CI)</td>
<td>ICCt1–t2 (F; CI)</td>
</tr>
<tr>
<td>bp ratio</td>
<td>0.47 (2.74; 0.18–0.68)**</td>
<td>0.23 (1.73; −0.04–0.53)*</td>
<td>0.62 (4.28; 0.39–0.78)**</td>
</tr>
<tr>
<td>pp ratio</td>
<td>−0.03 (0.93; −0.33–0.28)§</td>
<td>0.35 (2.05; 0.05–0.59)*</td>
<td>0.58 (3.72; 0.33–0.75)**</td>
</tr>
<tr>
<td>bp difference</td>
<td>0.74 (6.55; 0.55–0.85)**</td>
<td>0.7 (5.63; 0.50–0.83)**</td>
<td>0.79 (8.72; 0.64–0.89)**</td>
</tr>
<tr>
<td>pp difference</td>
<td>0.75 (6.97; 0.57–0.86)**</td>
<td>0.63 (4.47; 0.41–0.79)**</td>
<td>0.82 (9.93; 0.68–0.90)**</td>
</tr>
</tbody>
</table>

CI: 95% confidence interval. **p < 0.001, *p < 0.05, § not significant. Two-way mixed effects model with fixed measures effects and single measurements: Model (3.1). F-test with true value 0.
4. Discussion

Reproducibility is an essential requirement to disease markers, particularly when considered intermediate phenotypes. We evaluated retest reliability of auditory sensory gating – a possible intermediate phenotype – in P50, N100 and P200 i.e. sequentially occurring evoked components in a sample of 41 healthy subjects.

As indicated by their ratio- and difference-gating values, all three evoked components showed sensory gating. The strongest sensory gating was observed for the P50 component; nevertheless both N100 and P200 exhibited considerable sensory gating (ratio-gating: 35–38%, 48–54%, 45–57%, respectively). As indicated by their correlation values P50, N100 and P200 difference-gating clearly shared some variance with each other. This was not the case for gating measured as ratio scores.

Although we used shorter inter-pair intervals (2.8 s) than most P50 sensory gating studies (8 s), the sensory gating values were in line with those reported by others using similar amplitude measurement methods (Boutros et al., 2006; Kizkin et al., 2006; Waldo et al., 1988). We chose shorter intervals, because there is evidence for full P50 response recovery in the paired-stimuli design even at intervals as short as 2 s (Dolus et al., 2001). Recently, we found comparably P50 and N40–P50 ratio-gating values for short and long inter-pair intervals (Rentzsch et al., submitted for publication).

However, a broad range of ratio-gating values have been reported in healthy subjects even with the long inter-pair interval, e.g. P50 ratio-gating 16–50% (Adler et al., 1985; Boutros et al., 2004; Clementz et al., 1998a; Freedman et al., 1996; Kathmann and Engel, 1990; Olincy et al., 2000); N100 ratio-gating 24–58% (Adler et al., 1990b; Boutros et al., 1999b, 2006; Kisley et al., 2003; Waldo et al., 1988; Yee et al., 1998); and P200 ratio-gating 15–57% (Hetrick et al., 1996; Boutros et al., 2004, 2006). This may be related to additional differences in stimulus design, filter settings and amplitude-measurement methods.

A primary requirement for quality gating values is good amplitude reliability (Smith et al., 1994). We found good to excellent reliabilities for all amplitudes, except for the second N100 amplitude, for which reliability was poor.

Ratio-gating is the most common method for sensory gating estimation and is thus of particular interest. Briefly, our results showed good reliability for the P200 ratio-gating, but only fair reliability for the P50 ratio-gating. The N40–P200 ratio-gating reliability was fair whereas the N40–P50 ratio-gating reliability was only poor. For the ratios of the N100 and N40–N100, reliability was considered poor. Notably, for all difference-gating values reliability was good to excellent.

Method of amplitude measurement influenced only ratio-gating reliability of the P50 component, but not that of the N100 and P200 component. For P50 ratio-gating, i.e. the ratio of the P50 components referenced to pre-stimulus baseline, reliability was considered fair. In contrast, reliability for the classical N40–P50 ratio-gating, i.e. P50 component referenced to N40 component, was inadequate; this was in line with previous results (Boutros et al., 1991a,b; Clementz et al., 1998a; Hall et al., 2006; Kathmann and Engel, 1990; Lamberti et al., 1993; Naber et al., 1992; Smith et al., 1994). Surprisingly however, most clinical studies still use N40–P50 ratio-gating instead of P50 ratio-gating; e.g. investigation of cognitive deficits (Braff and Light, 2004); of central drug effects (Brunstein et al., 2005; Ghisolfi et al., 2002; Olincy et al., 2006); and of intermediate phenotypes (Cadenhead et al., 2002; Freedman et al., 2000). It is discussed that referring P50 to the N40 amplitude (N40–P50) is superior, because referring the P50 to pre-stimulus baseline is affected by: a) a contingent negative variation interfering with the pre-stimulus baseline (Smith et al., 1994) and b) by the subsequent N100 component affecting the P50 itself (Jerger et al., 1992; Nagamoto et al., 1989).

Yet, our results do not confirm this. We found a distinctly better P50 than N40–P50 ratio-gating reliability. This may be a consequence of the high-pass 10 Hz filter, which attenuates slow frequencies underlying the contingent negative variation and the N100 component (see Fig. 1). Furthermore, the N40–P50 amplitude may be influenced not only by a predominant P30 (Cardenas et al., 1993) but also by variations of the N40 component (Yoshiura et al., 1995).

Although both the N40 and the P50 amplitudes show good to excellent reliabilities, this was not observed for the N40–P50 ratio-gating. The reliability of the N40–P50 ratio-gating improved after post-hoc exclusion of subjects with a low first P50 amplitude. This reduced the sample size by 22%, thus questioning the usefulness of this procedure. Since schizophrenic patients show lower P50 amplitudes compared to healthy controls, this could lead to higher exclusion rates and to a possible population bias.

Several reasons for the previously mentioned low N40–P50 ratio-gating reliability have been pointed out. Given that both the first and second P50 amplitudes are not independent of each other, and that the shared variance cannot be eliminated completely, reliable ratio-gatings are difficult to obtain (Smith et al., 1994). Further, all cerebral evoked potential measurements include both a signal and a noise component. Thus, a ratio measurement that includes noise both in the numerator (second P50) and denominator (first P50) will introduce an even greater variability. Although this explains the higher reliability of the difference-gating compared to the ratio-gating, it does not explain the very low reliability of the N40–P50 ratio-gating compared to the P50 ratio-gating satisfactorily.

The test–retest reliabilities of difference-gating values were mostly good or excellent and similar for both amplitude-measurement methods. The superior reliability for the N40–P50 difference-gating compared to the N40–P50 ratio-gating is a well established finding (Hall et al., 2006; Jerger et al., 1992; Smith et al., 1994). It is therefore surprising, that the difference-gating method plays only a marginal role in P50 sensory gating studies, in contrast to the ratio-gating method. This is the case, despite P50 difference-gating also being a useful tool to investigate sensory gating deficits in schizophrenic patients (Becker et al., 2004; Clementz et al., 1998a,b) and P50 sensory gating being described in the literature in a manner that is fully compatible with difference-gating as sensory gating index (Smith et al., 1994).
The test–retest reliabilities for the P200 ratio-gating were fair to good and for the N100 ratio-gating poor, both independently of amplitude-measurement method. Hence, P200 ratio-gating is stable over time, whereas N100 ratio-gating may not. The dependence of the N100 amplitudes on arousal and attention may be an explanation for the low reliability of its ratio-gating (Cardenas et al., 1997; Erwin and Buchwald, 1986; Gallinat et al., 2002; Jerger et al., 1992; White and Yee, 1997). Interestingly, the P200 component shows an opposite dependence on arousal and attention to that of the N100; e.g. the amplitude of the P200 decreases with attention, shows no habituation and increases at sleep onset (for review see: Crowley and Colrain, 2004). This may explain the better reliability of the P200 sensory gating.

We note that the reliabilities of the different sensory gating indices do not necessarily apply to clinical samples. For instance, the short inter-pair interval used here may have a different influence on the sensory gating in schizophrenic patients and thus result in a different long-term stability compared to healthy subjects. However, demonstrating adequate long-term stability in healthy subjects is a precondition when investigating clinical samples.

In a recent twin study, Anokhin et al. (2006) found for the classical peak-to-peak P50 ratio-gating only a modest heritability but a high heritability for the N100 and P200 ratio-gatings, as well as for the P50, N100 and P200 difference-gating. According to this, we conclude that the P200 ratio-gating and the difference-gating of all three responses may well serve as stable phenotypes, at least in healthy subjects.

Acknowledgements

Nash N. Boutros was supported by grant: RO1-MH58784. We are grateful to Dr. Hans Dorn for his technical assistance.

References


