
The effect of measurement error on the test–retest reliability of repeated mismatch negativity measurements

Antti K.O. Paukkunen, Miika Leminen, Raimo Sepponen

Electronics Department, School of Science and Technology, Aalto University, 00076 Aalto, Finland
Graduate School of Electrical and Communications Engineering, School of Science and Technology, Aalto University, 00076 Aalto, Finland
Institute of Behavioral Sciences, Helsinki University, 00140 University of Helsinki, Finland
Finnish Center of Excellence in Interdisciplinary Music Research, Finland

Objective: The aim was to study how the measurement error affects the repeatability of mismatch negativity (MMN) measurements.

Methods: Event-related potentials (ERPs) to changes in sound frequency, location, intensity, duration, and composition were recorded five times during 1–3 weeks from 13 healthy adults using a multi-feature MMN paradigm. The accumulation of MMN was modeled empirically with respect to measurement error, and repeatability was estimated at 0.6–3.5-μV error levels. The analysis was made for the results in the single deviant conditions and their pattern (auditory discrimination profile).

Results: At the single-subject level, the measurement error significantly affected the repeatability until it went below 9–17% of MMN peak amplitude. At the group level, the threshold was higher. Peak amplitude was generally the most repeatable parameter. Latency was superior when the error was moderate or small (<2–3 μV).

Conclusions: The measurement error affects the repeatability of MMN. In single-subject studies, it should not be neglected if it exceeds 10% of the MMN amplitude. The application of the auditory discrimination profile is recommended for future applications.

Significance: The study provided quantitative results to support the discussion on improving the repeatability of the MMN measurements. They are expected to apply conditionally to other ERP measurements, too.

1. Introduction

During the last few decades, an increasing number of studies have demonstrated the application of auditory event-related potentials (AERPs) in the investigation of neurological and psychiatric disorders. One of the most popular measures appearing in these studies is mismatch negativity (MMN, Näätänen et al., 1978). MMN indicates the discrimination of a change in the sensory input (Näätänen, 1990, 1992; Näätänen and Winkler, 1999; Näätänen et al., 2001; Garrido et al., 2009) and it can be measured by analyzing the difference in the AERPs elicited by slightly different stimuli (Schröger, 1998; Duncan et al., 2009). Applications of MMN include the investigation of, e.g., dyslexia, schizophrenia, memory disorders, and coma outcome (for reviews see Kujala et al., 2007; Näätänen et al., 2007; Duncan et al., 2009). So far, the focus has been on general research, but MMN is also a tempting option for clinical use because of its versatile nature.

However, like many other AERP components (which are modulated by cognitive functions of the brain), MMN is not yet suited to clinical applications, because it cannot be measured reliably enough (Kujala et al., 2007). According to the reported studies, the test–retest reliability of MMN varies between 0.37 and 0.87, depending on the experimental conditions (Chertoff et al., 1988;
2.1. Test procedure

Thirteen healthy volunteer test subjects (9 males, age: 20–28) participated in the study by attending a one-hour MMN recording session which was repeated five times during a period of 1–3 weeks. The recording time would be too long for practical use of MMN, the single-subject measurements need to be repeatable, too (Kujala et al., 2007; Näätänen et al., 2007).

The first systematic studies that aimed at improving the test–retest reliability of MMN were reported by Pekkonen et al. (1995) and Lang et al. (1995). It was found that the repeatability depends on the presentation of the stimuli, the number of trials recorded (Pekkonen et al., 1995), the parameterization of MMN, and the characteristics (e.g., age, discrimination ability, and alertness) of the test subjects (Lang et al., 1995). This provided a good starting point for the research and still forms the basis of the methods that are used today. Currently, the best way to secure reliable recordings is to take care of the signal quality (Piviä et al., 1993; Sinkkonen and Tervaniemi, 2000), to use appropriate stimuli (Duncan et al., 2009), and to apply efficient recording procedures, such as the multi-feature paradigm (Näätänen et al., 2004; Pakarinen et al., 2007, 2009). Valid stimulus design and a high signal-to-noise ratio (SNR) reduce the variation in the results (Lang et al., 1995; Sinkkonen and Tervaniemi, 2000). An efficient recording procedure, on the other hand, permits efficient denoising through averaging and robust rejection of contaminated data while not causing unnecessary mental fatigue and stress for the test subject (e.g., Pakarinen et al., 2009). In addition, distracting the test subjects from attending to the stimuli also improves the repeatability as a result of the reduced modulation of the recorded responses (Näätänen, 1995).

Considering the further development, many authors have suggested that a major part of the uncertainty would be contributed by the measurement error (e.g., Lang et al., 1995; Sinkkonen and Tervaniemi, 2000; Hall et al., 2006; Paukkunen et al., 2010a). As a smaller error yields higher repeatability, it is probable that even a small error could have a major effect on the repeatability of a weak response like MMN. Quantitative studies, however, have not been published on the subject and the extent to which the effect is relevant has not been properly evaluated. In this study, a series of repeated MMN recordings is made with multiple test subjects to analyze the effect of the measurement error in vivo.

The main objective is to determine how the test–retest reliability of the results changes as a function of the measurement error. In addition, it is studied how this affects the parameterization of MMN.

2.2. Stimulus presentation

Six different types of stimuli were used (a standard tone and five deviants), and they were presented through stereo headphones. The standard tone (Std), the frequency deviant (Freq), the location deviant (Loc), and the duration deviant (Dur) were similar to the ones used in Pakarinen et al. (2007). The intensity deviant (Int) and the gap deviant (Gap) were implemented according to the recommendations given in Duncan et al. (2009).

The standard tone was a 75-ms-long (including 5-ms rise and fall times) harmonic sinusoidal tone with a 523-Hz fundamental frequency and two harmonics (1046 Hz and 1569 Hz). It was presented at an intensity of +40 dB above the individual hearing threshold. The deviant stimuli deviated from the standard tone in the fundamental frequency (Freq, low: 450 Hz, or high: 609 Hz), the perceived location of the sound source (Loc, ±90°), intensity (Int, ±10 dB), duration (Dur, −48 ms), and temporal composition (Gap, 5–ms silent period in the middle of the tone). The location deviant was produced by introducing a 700-µs inter-aural difference between the left and right audio channels.

The responses to all stimulus types were obtained in each session, and they were elicited by using a multi-feature paradigm (Näätänen et al., 2004; Pakarinen et al., 2007, 2009), where every other stimulus is a standard tone followed by a deviant tone. The total number of stimuli presented during each session was 7200, half of which were standard tones (p = 0.5). The deviant stimuli were presented equiprobably (p = 0.1) so that the two sides of the two-sided deviants (Freq, Loc, Int) were equiprobable (p = 0.05). The stimulus onset asynchrony (SOA) was 500 ms.

2.3. AERP measurement

During the recording session, the test subjects watched a neutral movie with subtitles and no sounds. The EEG was recorded (0.17–70 Hz, 200-Hz sampling rate) from eight channels (F7, F8, Fp1, Fp2, Cz, Fz, and the mastoids (M2, M2)) by using a passive electrode cap (Ag/AgCl electrodes, attached with electrode gel) and an 8-channel EEG amplifier (Paukkunen et al., 2010b) with common average reference. After digitalization, the data recorded were first inspected to detect artifacts, re-referenced to the mean of the mastoids and filtered by using a digital 30-Hz lowpass filter. Then they were divided into epochs of 350 ms (−50 ms to +300 ms from the stimulus onset) and the baseline was corrected by subtracting the mean of the 50-ms prestimulus interval.

In the artifact analysis, no particular electro-oculogram (EOG) channel was used, but the eye blinks were detected from the frontal electrodes Fp1 and Fp2. The amplitude of the blinks typically varied from 40 µV to 80 µV, depending on the test subject, and prior to averaging, the epochs with amplitude in excess of ±20 µV (at Fp1 or Fp2) were rejected. On the basis of a visual inspection of individual epochs, the rejection criterion was found to be appropriate. The rejection ratio was high and 36 ± 16% (mean ± sd, standard deviation) of the data was rejected because of suspected artifacts. This, however, was not considered to be a problem, since the number of epochs recorded was also large.
2.4. Data analysis

2.4.1. Overview

The analysis was started by averaging the accepted responses to each stimulus type and calculating the respective measurement error and MMN parameters as a function of the number of epochs included into the sum \(N\). Then, the data from the first part of the analysis were joined to present the MMN parameters as a function of the measurement error, and to study the effect of the error on the test–retest reliability of the parameters.

2.4.2. Quantification of the MMN

MMN (Fig. 1A) was derived by subtracting the responses to the standard tones from those to each of the deviants and the magnitude was parameterized by using three alternative estimators: peak amplitude, mean amplitude, and latency. The peak amplitude (PEAK) and latency (LAT) were estimated by determining the minimum of the difference waveform at an interval of 100–200 ms (from the stimulus onset). The mean amplitude (MEAN) was estimated by calculating the mean amplitude of the signal in a fixed 100-ms time window that was centered at 150 ms (from the stimulus onset).

2.4.3. Quantification of the measurement error

The measurement error (Fig. 1B) was determined on the basis of the (+/−) average (Schimmel, 1967), i.e., the difference between the averages of the odd and the even epochs included in the sum. Based on the cancellation of the event-locked responses in the subtraction process, it allowed the estimation of the residual noise in the average waveform (Schimmel, 1967; Ruchkin, 1988), and the determination of the error could be made on the basis of the statistics of the noise waveform (e.g., Lowy and Weiss, 1968; Wong and Bickford, 1980; Elberling and Don, 1984; van de Velde et al., 1996; Gerull et al., 1996). First, the (+/−) average was computed and the error estimates quantified separately for each stimulus condition. Then the errors estimated for the standard and deviant responses were summed to produce an error estimate for the MMN components, in each deviant condition. As the sign of the terms was unknown and they could not be assumed to be positively correlated, summing was necessary to avoid underestimating the total effect.

As the level of interference is neither topographically uniform through the EEG channels nor temporally stationary through the epochs (Picton, 1995; Gerull et al., 1996), the error estimations were not expected to apply outside the analysis time frame. Thus, to make them specific to MMN, the (+/−) average was computed in the same time window (100–200 ms from the stimulus onset) and EEG channel (Fz, re-reference to the mean of mastoids) used to determine the MMN parameters. Furthermore, on the basis of the (+/−) averages, the magnitude of the error was quantified by determining the average deviation over the time window, and dividing it by 2, according to Eq. (1):

\[
\text{ERR} = \frac{1}{2n} \sum_{t=100}^{200} |\mu_{es}[t] - \mu_{os}[t]| + \frac{1}{2n} \sum_{t=100}^{200} |\mu_{ed}[t] - \mu_{od}[t]| \tag{1}
\]

where \(n\) represents the number of samples at the interval [100, 200] ms, and \(\mu_{es}[t], \mu_{os}[t], \mu_{ed}[t], \text{ and } \mu_{od}[t]\) the partial average waveforms calculated for the even (e) and odd (o) epochs, and standard (s) and deviant (d) data, respectively.

With respect to the MEAN component, this gave a direct estimation of the error. It was also relevant to PEAK and LAT, because the goodness of the responses had an effect on the localization of the maximum peak and thus the accuracy of the estimated components. Alternatively, the error could also have been quantified on the basis of the single-point deviation of the (+/−) average, at the peak latency of MMN, or by determining the maximum deviation at the analysis interval. The pointwise estimator would have provided more specific information on the error of the MMN peak, but it became distorted when the peak latency could not be accurately located. The maximum estimator, on the other hand, would have given a realistic maximum estimate for the error. It was,
however, upward-biased for most of the MMN waveform, thus making it not specific enough to the parameters being studied.

2.4.4. Estimation of the test–retest reliability

After the MMN and the measurement error as a function of N had been estimated, the information was joined to create an empirical model of the progression of the MMN parameters as a function of the measurement error. This part of the analysis was no longer performed for all the data, but to maximize the quality, the computations were only made for the data extracted from Fz, where the signal quality was generally the highest (Fig. 2).

First, the MMN values were grouped with respect to measurement error at 29 different levels (0.6–3.5 µV, evenly distributed) (Fig. 3A, C and E). Then, for each level, a representative value was

Fig. 2. Distribution of the mean MMN amplitude (MEAN), and average signal-to-noise ratio (SNR; c.f. Möcks et al., 1988), averaged over the test subjects at the time interval of 100–200 ms from the stimulus onset. The signal quality was generally the highest at Fz.

Fig. 3. Progression of the MMN peak amplitude (A and B), MMN mean amplitude (C and D), and MMN latency (E and F) as a function of the measurement error in the five repetitive recording sessions, in the example case (subject 13, frequency deviant). The original grouped data are presented in the first column (A, C, E) and the smoothed data in the second column (B, D, F). The data from the different sessions (T1–T5) are indicated by different colors and symbols.
determined by calculating the mean of the samples in a 0.2-µV window, and the missing values were linearly interpolated according to their nearest neighbors (Fig. 3B, D and F). The operation allowed a continuous model of the progression of MMN to be created. This was necessary because of the irregular distribution of the measurement error values that occurred at the interval studied. It also allowed most of the outliers from the data to be rejected.

The models were created separately for each session, and deviant condition. Then, on the basis of them, the test–retest reliability was studied by calculating the Pearson’s correlation coefficient. First, the analysis was performed separately for the result in the different deviant conditions (Freq, Loc, Int, Dur, and Gap). Then it was performed for the auditory discrimination profile (Pakarinen et al., 2007), which included responses for all the different deviant conditions. In the investigation of the profile, the analysis was conducted for both the individual results and the group average. In the investigation of the responses to the different stimulus types, it was made at the single-subject level only.

3. Results

3.1. AERPs and MMN

The grand average of the responses recorded for each type of stimulus, the respective MMN parameters, and their variation across the recording sessions are presented in Fig. 4. The MMN peak amplitude was the highest for Freq (PEAK: $-3.0 \pm 1.1$ µV, MEAN: $-1.9 \pm 0.9$ µV) and Dur (PEAK: $-3.0 \pm 0.9$ µV, MEAN: $-2.0 \pm 0.7$ µV). The smallest response was produced by Loc (PEAK: $-1.8 \pm 0.8$ µV, MEAN: $-1.0 \pm 0.7$ µV). The peak latency of the response was the longest for Int (LAT: 177 ms ± 23 ms), while Loc had the shortest peak latency (LAT: 121 ms ± 24 ms).

3.2. Accumulation of the average responses

Fig. 5 shows representative results (subject 13) that demonstrate how the variation in the differential responses changes when the measurement error decreases. In the example shown, the variation in the MMN parameters between the repeated measurements was reduced for all the deviant conditions at the interval that was studied (1–3 µV). The reduction in the range of the observations was about 1.1–5.3 µV (average: 3.5 µV, $p < 0.05$, paired t-test) at the peak amplitude, 1.4–5.9 µV (average: 4.0 µV, $p < 0.05$) at the mean amplitude, and 20–55 ms (average: 33 ms, $p < 0.05$) at the peak latency, depending on the stimulus type.

3.3. Test–retest reliability

3.3.1. Single deviant responses

According to the results from the investigation of MMN in the single deviant conditions (Freq, Loc, Int, Dur, and Gap), the test–retest reliability of the responses was generally improved with the decreasing measurement error (Fig. 6). From the error level of 3–1 µV, PEAK repeatability was improved by 0.29–0.39 ($p < 0.05$, paired t-test), and MEAN repeatability by 0.29–0.35 ($p < 0.05$). The repeatability of LAT, on the other hand, was significantly increased only in the Loc and Dur conditions ($\Delta = 0.40–0.71$, $p < 0.05$), while the change was insignificant in the other conditions ($\Delta = -0.06–0.16$, $p = 0.18–0.55$).

In most deviant conditions, the effect was insignificant or small at the higher error levels (>2 µV), but it became higher when the error went below the MMN amplitude. From the error level of 3–2 µV, the change in the PEAK repeatability was significant ($p < 0.05$, paired t-test) only in the Int and Dur conditions ($\Delta = 0.18–0.21$), the change in the MEAN repeatability only in the Freq and Dur conditions

$\Delta = -0.08–0.12$, and the change in the LAT repeatability only in the Freq, Loc, and Int conditions ($\Delta = -0.39–0.2$). On the other hand, from the error level of 2–1 µV, the change in the PEAK repeatability was significant in the Freq, Loc, Dur, and Gap conditions ($\Delta = 0.19–0.48$), the change in the MEAN repeatability in all the conditions ($\Delta = 0.23–0.37$), and the change in the LAT repeatability in the Freq, Loc, and Dur conditions ($\Delta = 0.23–0.51$).

The maximum repeatability reached was 0.324–0.743/0.391–0.676/0.158–0.678 (PEAK/MEAN/LAT), depending on the stimulus condition. This, however, did not represent the ultimate maximum, and most of the curves still had a clear increasing trend at the smallest error level studied (0.6 µV).

3.3.2. Auditory discrimination profile

Like the repeatability of MMN in the single deviant conditions, also the test–retest reliability of the auditory discrimination profile also increased with a decreasing error level at the error interval studied (Fig. 7). From the error level of 3–1 µV, it was improved by 0.43/0.30/0.45 (PEAK/MEAN/LAT, $p < 0.05$) on average at the single-subject level (single subjects) and by 0.31/0.16/0.37 (PEAK/MEAN/LAT, $p < 0.05$) at the group level (group average). The maximum repeatability was 0.712/0.556/0.691 (PEAK/MEAN/LAT), on average, at the single-subject level, and 0.942/0.904/0.987 (PEAK/MEAN/LAT) at the group level. On average, the
Fig. 5. The difference of the response to the deviant and standard tones (i.e., MMN) at three error levels (3 μV, 2 μV, and 1 μV) in the five repeated sessions, in the example case (subject 13). Responses to the different types of stimuli (Freq, Loc, Int, Dur, and Gap) are presented in columns 1–5, and the results attained with different error levels (Err) are presented in rows 1–3. The repeated measurements (T1–T5) are indicated by different line styles.

Fig. 6. The estimated test–retest reliability (Pearson’s r) of the MMN responses elicited in the different deviant conditions (Freq, Loc, Int, Dur, Gap) as a function of the measurement error (± average, error). The estimates were computed on the basis of (A) MMN peak amplitude (PEAK), (B) MMN mean amplitude (MEAN), and (C) MMN latency (LAT), and averaged over all the subjects.

Fig. 7. The estimated test–retest reliability (Pearson’s r) of the auditory discrimination profile (Pakarinen et al., 2007) as a function of the measurement error (± average, error). (A) MMN peak amplitude (PEAK), (B) MMN mean amplitude (MEAN), and (C) MMN latency (LAT). The solid line represents the average test–retest reliability in single-subject measurements and the dashed line represents the test–retest reliability of the group average.
repeatability was approximately 46% better at the group level than the single-subject level. At the group level, the effect of the error was clear when the error was higher than 1–2 µV, but it was decreased at the smaller levels, and had only a small effect when the error was less than 1/0.7/1.3 µV (PEAK/MEAN/LAT). At the single-subject level, on the other hand, the curves still had a clear rising trend at the smallest error level studied (0.6 µV), and a similar threshold could not be identified directly from the data. However, by considering that the effective error was about 3–4 times higher than the error in the group profiles as a result of the averaging effect (Möcks et al., 1988; Regan, 1989; Lang et al., 1995), it could be estimated that the effect would have been of only small relevance if the error had gone below about 0.3/0.2/0.4 µV (PEAK/MEAN/LAT).

4. Discussion

First, the results of the present study show that the effect of the measurement error on the test–retest reliability of MMN is dominant at the higher error levels, but it decreases with a smaller error. At the single-subject level, the effect was clear when the error was on the same scale as the MMN, and it was found to be increasing at least until the error went below 0.6 µV (20–33% of the MMN peak amplitudes, mean: 26%). The level where the measurement error would become irrelevant could not be determined directly from the data. However, on the basis of the analysis of the group averages that was performed, it could be estimated that an error below 0.2–0.4 µV (7–22% of the MMN peak amplitudes, mean: 9–17%) would only have a small effect on the repeatability of the responses. The estimations were conducted from the results related to the analysis of the auditory discrimination profile. However, they are also expected to apply to the single deviant responses, because the effect of the error on the repeatability (see single-subject data in Figs. 6 and 7) was quite similar in both cases ($r = 0.66–0.96$, $p < 0.05$).

Second, it was found that at the single-subject level, the test–retest reliability of the auditory discrimination profile was superior to that of the single deviants. The maximum repeatability of the discrimination profile was 0.712/0.536/0.691 (PEAK/MEAN/LAT), and the maximum repeatability of the single deviant types was 0.570/0.513/0.454 (PEAK/MEAN/LAT) on average. The application of the discrimination profile is based on an inter-comparison of the magnitude of the parameters in different deviant conditions. This allows the values to be threaded as being normalized. Therefore, the profile is tolerant of systematic changes in the magnitude of the parameters and it has a positive effect on the repeatability. In addition, it may also be that the comparative analysis method handles the "no response" condition better, for it provides a reference for the evaluation of the relevance of the magnitude of the response. This information is not available in the investigation of the responses to single deviants and the rating of the results may be complicated.

Third, the measurement error had a different effect on the different MMN parameters (PEAK, MEAN, LAT). According to the results, PEAK was the most stable parameter in the overall performance. The maximum test–retest reliability of LAT was at the same level as PEAK, but it was unstable at higher error levels, probably because it was hard to quantify the latency when the SNR was low. The MMN peak latency is generally considered to be more reliable than the amplitude (e.g., Lang et al., 1995). However, the present results (see group data in Fig. 7) suggest that the assumption only holds true when the measurement error is moderate or small (less than 2–3 µV). At the higher error levels, both PEAK and MEAN proved to be at least as good. In addition, it should be noted that the investigation of the amplitudes probably allows better consideration of the "no response" situation because they always converge towards zero. The latency estimate only gives random results when no clear response is present.

In conclusion, the results of the present study showed that the measurement error affects the repeatability of the repeated MMN measurements independently of the parameterization and the way the results are analyzed. At the single-subject level, it should not be neglected until the error goes at least below 10% of the MMN peak amplitude. This is a very stringent requirement and, since the quality is only improved relative to the square root of the number of trials averaged, optimization of the signal quality is critical and data processing has to be made carefully. At the group level, the criteria may be adjusted according to the size of the test group. For example, regarding the auditory discrimination profile, averaging of 150 trials left an error of about 0.5–1.5 µV, which allowed a repeatability of 0.8–0.98, at the group level, and 0.3–0.7, at the single-subject level.

Although the tests were performed for the MMN, they conditionally apply to other evoked potential measurements, too. The exact criteria presented will not necessarily hold true in general and the behavior of the parameters depends on the type of response being studied. The effect of the measurement error, however, is similar provided that the responses that are studied also remain stable within the recording session. Furthermore, the method applied for estimating the error will also have an effect on the results, and the quantification of the error has to be defined with respect to the parameters studied.

Acknowledgements

The work was supported in part by the Graduate School of Electrical and Communications Engineering, Society of Electronics Engineers, and the Academy of Finland (National Centers of Excellence 2006–2011). The authors would like to thank Riitta Hari, Risto Näätänen, and Pekka Eskelinen for their assistance in the preparation of the study and discussions during the preparation of the manuscript.

References
