HHS Public Access

Author manuscript

Clin Neurophysiol. Author manuscript; available in PMC 2021 December 01.

Published in final edited form as:

Clin Neurophysiol. 2020 December ; 131(12): 2899–2909. doi:10.1016/j.clinph.2020.09.027.

Reliability of Mismatch Negativity Event-Related Potentials in a Multisite, Traveling Subjects Study

Brian J. Roach[1,*], **Ricardo E. Carrión**[2,3,4,*], **Holly K. Hamilton**[1,5], **Peter Bachman**[6], **Aysenil Belger**[7], **Erica Duncan**[8,9], **Jason Johannesen**[10], **Gregory A. Light**[11,12], **Margaret Niznikiewicz**[13], **Jean Addington**[14], **Carrie E. Bearden**[15], **Kristin S. Cadenhead**[11], **Tyrone D. Cannon**[10,16], **Barbara A. Cornblatt**[2,3,4,17], **Thomas H. McGlashan**[10], **Diana O. Perkins**[7], **Larry Seidman**[13], **Ming Tsuang**[11], **Elaine F. Walker**[18], **Scott W. Woods**[10], **Daniel H. Mathalon**[1,5]

[1.]San Francisco Veterans Affairs Healthcare System, San Francisco, CA, United States

[2.]Division of Psychiatry Research, The Zucker Hillside Hospital, North Shore-Long Island Jewish Health System, Glen Oaks, NY, United States

[3.]Center for Psychiatric Neuroscience, Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, NY, United States

[4.]Department of Psychiatry, Hofstra North Shore-LIJ School of Medicine, Hempstead, NY, United States

[5.]Department of Psychiatry, University of California, San Francisco, CA, United States

[6.]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, United States

[7.]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

[8.]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA, United States

[9.]Atlanta Veterans Affairs Medical Center, Decatur, GA, United States

[10.]Department of Psychiatry, Yale University, School of Medicine, New Haven, CT, United States

[11.]Department of Psychiatry, University of California, San Diego, La Jolla, CA, United States

[12.]Veterans Affairs San Diego Healthcare System, La Jolla, CA, United States

Correspondence: Daniel H. Mathalon, Ph.D., M.D., San Francisco VA Healthcare System/ Psychiatry Service (116D), 4150 Clement Street, San Francisco, CA 94121, USA, Tel.: +1-415-221-4810, x 23860, daniel.mathalon@ucsf.edu.

*Indicates shared first authorship.

Conflict of Interest

None.

[13]Department of Psychiatry, Harvard Medical School at Beth Israel Deaconess Medical Center and Massachusetts General Hospital, Boston, MA, United States

[14]Hotchkiss Brain Institute, Department of Psychiatry, University of Calgary, Calgary, Alberta, Canada

[15]Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, CA, United States

[16]Department of Psychology, Yale University, School ofMedicine, New Haven, CT, United States

[17]Department of Molecular Medicine, Hofstra North Shore-LIJ School of Medicine, Hempstead, NY, United States

[18]Department of Psychology, Emory University, Atlanta, GA, United States

Abstract

Objective.—To determine the optimal methods for measuring mismatch negativity (MMN), an auditory event-related potential (ERP), and quantify sources of MMN variance in a multisite setting.

Methods.—Reliability of frequency, duration, and double (frequency+duration) MMN was determined from eight traveling subjects, tested on two occasions at eight laboratory sites. Deviant-specific variance components were estimated for MMN peak amplitude and latency measures using different ERP processing methods. Generalizability (G) coefficients were calculated using two-facet (site and occasion), fully-crossed models and single-facet (occasion) models within each laboratory to assess MMN reliability.

Results.—G-coefficients calculated from two-facet models indicated fair (0.4<G<=0.6) duration MMN reliability at electrode Fz, but poor (G<0.4) double and frequency MMN reliability. Single-facet G-coefficients averaged across laboratory resulted in improved reliability (G>0.5). MMN amplitude reliability was greater than latency reliability, and reliability with mastoid referencing significantly outperformed nose-referencing.

Conclusions.—EEG preprocessing methods have an impact on the reliability of MMN amplitude. Within site MMN reliability can be excellent, consistent with prior single site studies.

Significance.—With standardized data collection and ERP processing, MMN can be reliably obtained in multisite studies, providing larger samples sizeswithin rare patient groups.

Keywords

Event-Related Potentials (ERP); Mismatch Negativity (MMN); EEG; Reliability; Intra-class correlation coefficients (ICC); Psychosis

Introduction

Mismatch negativity (MMN) is an event-related potential (ERP) component that is automatically elicited by an infrequent deviant auditory stimulus that differs in pitch, duration, or another sound feature from a repetitive series of preceding “standard” stimuli. MMN is considered to reflect sensory echoic memory, since the detection of auditory

deviance depends on the short-term online formation of a memory trace of the immediately preceding standard sounds in the auditory processing stream, and can be measured using either electroencephalography (EEG) or magnetoencephalography (MEG). The MMN has great potential as an ERP biomarker because of its robust sensitivity to the pathophysiology of schizophrenia (Avissar et al., 2018, Erickson et al., 2016, Umbricht and Krljes, 2005), sensitivity to subclinical psychotic symptoms in the general population (Doring et al., 2016), and its ability to predict transition to a psychotic disorder in individuals at clinical high-risk (CHR) (Bodatsch et al., 2011, Perez et al., 2014, Shaikh et al., 2012). In addition to psychosis and psychosis-risk research, MMN has a broad scope of potential clinical applications and has been used extensively to study other disorders, including but not limited to comatose/disorders of consciousness (Andre-Obadia et al., 2018), Alzheimer's disease (Danjou et al., 2019, Horvath et al., 2018), bipolar disorder (Hermens et al., 2018), specific language impairment (Kujala and Leminen, 2017), substance use disorders (Ramlakhan et al., 2018), autism (Schwartz et al., 2018), and dyslexia (Volkmer and Schulte-Korne, 2018). While the test-retest reliability of MMN has been the focus of several studies within a single laboratory, the reliability and consistency of the MMN response across testing location must be evaluated in order to determine the suitability of this ERP component for use in multi-site, clinical trials or longitudinal studies of clinical populations.

Many test-retest reliability studies of MMN relied upon Pearson (Kathmann et al., 1999, Kujala et al., 2001, Pekkonen et al., 1995, Schroger et al., 2000, Tervaniemi et al., 1999, Uwer and von Suchodoletz, 2000) or Spearman (Deouell and Bentin, 1998, Schall et al., 1999) correlation coefficients. Such coefficients only evaluate the degree to which MMN responses or ranks from two tests covary, without considering whether responses are in close agreement from one test occasion to the next. A better measure of such agreement is the intra-class correlation (ICC) coefficient (Shrout and Fleiss, 1979). Eight studies (Biagianti et al., 2017, Chen et al., 2018, Hall et al., 2006, Lew et al., 2007, Light and Braff, 2005, Light et al., 2012, McCleery et al., 2019, Recasens and Uhlhaas, 2017) have reported ICCs of MMN. In general, these studies have found that MMN reponses are stable over time, further highlighting the potential for the component to be used a biomarker in clinical populations (Naatanen et al., 2015). Regardless of the reported coefficient type, the majority of the above mentioned studies have only evaluated or compared the impact of different paradigms or sound features used to elicit the MMN. To our knowledge, no study has assessed the influence of different EEG signal processing choices, such as the reference electrode or methods of artifact rejection, on the reliability of the MMN responses despite the fact that these choices differ across reports.

In multi-site studies of the reliability of functional magnetic resonance imaging (fMRI) data, generalizability (G) theory has been applied to facilitate descriptions of the different sources of variance in blood oxygen level-dependent (BOLD) signal measurements(Brown et al., 2011, Forsyth et al., 2014). G-theory applies a random effects modeling approach to partition sources of variance by calculating variance components for the effect of persons as well as other measurement factors, or facets (e.g., study site, testing occasion), and their interactions. The present study focuses on a two facet, fully-crossed traveling subjects G-study of MMN from the North American Prodrome Longitudinal Study (NAPLS), which is a multi-site research consortium studying the mechanisms and predictors associated with

psychosis onset (Addington et al., 2012). The two facets were Site (NAPLS geographic location) and Occasion (test and retest day). This design is identical to previously published physiology (Cadenhead et al., 2013), fMRI (Forsyth et al., 2014, Noble et al., 2017) and structural MRI (Cannon et al., 2014) reports from the NAPLS-2 consortium because the same traveling subjects underwent EEG and MRI (for more NAPLS-2 details please see (Addington et al., 2012)) assessments on each test day.

The main goal of this study was to quantify variance components and associated G-coefficients representing the single site, single session reliability of the MMN response measured for clinical comparisons in NAPLS-2. We also present G-coefficients calculated with alternative methods of ERP averaging, MMN scoring, and referencing to numerically compare with the NAPLS-2 approach. Lastly, since previous MMN reliability studies were conducted at a single laboratory site, individual laboratory site G-coefficients were also calculated separately for each of the 8 NAPLS-2 sites and a “home” site model to allow for qualitative comparisons within the NAPLS consortium and the extant literature. While there are no hypothesis tests associated with G-coefficients, we expected MMN reliability using our NAPLS-2 measurement approach to be moderate to excellent (i.e., G-coefficient > 0.6), consistent with prior single site test-retest reliability reports. We expected MMN reliability calculated with the NAPLS-2 approach to be numerically equivalent and possibly greater than MMN reliability with alternative processing methods. We expected no difference in reliability between NAPLS-2 sites.

Methods

Participants

Traveling Subjects Sample—One healthy participant was recruited from each of the eight NAPLS-2 sites. Participants were excluded if they met criteria for a psychiatric disorder based on the Structured Clinical Interview for DSM-IV (First, 1997) or prodromal criteria based on the Structured Interview for Prodromal Syndromes (McGlashan et al., 2010, Miller et al., 2002), met criteria for substance dependence in the past 6 months, had a first-degree relative with a psychotic disorder, or had a neurological disorder. EEG data were collected on two consecutive test days at each site, starting at each participant’s home site followed by a pseudo-random travel order to all other sites. The average number of days between the first test occasion at each site was 7.3 (SD=7.6), and participants completed all 16 EEG sessions in 29 to 80 days. Participants were between 19 and 31 years old (mean=27.74, SD=3.99), and there were an equal number of males and females. All participants provided written informed consent and the protocol was approved by the Institutional Review Boards at each of the NAPLS sites.

Equipment

EEG—All participating data collection sites used BioSemi (www.biosemi.com) EEG acquisition systems. Half (UCLA, Harvard, UNC, Yale) of these systems were equipped to record 64channels of EEG, and half (Emory, Hillside, UCSD, Calgary) were equipped to record 32 channels which were located on standard, equidistant locations according to the international 10-20 system (Klem et al., 1999).

Stimulus Presentation—All sites used Dell Optiplex Desktop computers to run the MMN task (described below). These systems were configured to meet or exceed the minimum hardware requirements recommended at the time of study launch (2009) by the stimulus presentation software provider, neurobehavioral systems (www.neurobs.com), with special attention paid to video and sound cards. LCD monitors connected to a 512MB ATI Radeon PCIe video card by VGA cables were used at each site. Auditory stimuli were delivered via ER1 Etymotic insert earphones connected to a SoundBlaster X-Fi Xtreme Gamer PCI card. Subject responses were recorded with a Cedrus RB-830.

Paradigms

Hearing Test—Auditory stimuli were presented through ER1 Etymotic insert earphones using Presentation software. Prior to the MMN task, hearing levels were also assessed using the same stimulus presentation software and hardware employed in the MMN task. The hearing thresholds for three pure tones (500, 633, and 1000 Hz) were detected separately for each ear at the beginning of every session. This was accomplished by playing 50ms duration tones of each frequency in each ear, manipulating the “attenuation” parameter within the software. The attenuation value was set between 0 (no attenuation) and 1 (total attenuation). Starting with an attenuation value of 1, the value was decreased in 0.05 increments until the subject indicated that she had heard a tone in the target ear by pressing a left or right response button corresponding to the ear in which the tone was detected. A 0.05 step in attenuation is theoretically equivalent to 5 dB, but in practice the actual change in dB depends on the auditory stimulus delivery device and its frequency response function. Before starting the test, subjects were told to respond to the tones played through the right or left ear insert, and that the tones would never be played in both ears at the same time. After the subject’s first response to a specific tone, the attenuation value was increased by 0.2 or set to 1, whichever was less, and the process was repeated until the subject detected a tone from each frequency in each ear four times.

MMN Paradigm—Auditory stimuli delivery consisted of 85% standard tones presented for 50 ms at 633 Hz, 5% duration (DUR) deviants presented for 100 ms at 633 Hz, 5% frequency (FRQ) deviants presented for 50 ms at 1000 Hz, and 5% double-deviants (DBL) presented for 100 ms at 1000 Hz. A total of 1794 tones were presented over 3 separate blocks, with each block lasting approximately 5 minutes. Tones were presented with 5 ms rise and fall times and a 500 ms stimulus onset asynchrony. In an effort to reduce the effect of attention on the MMN ERPs, participants were instructed to ignore auditory stimuli while focusing on a separate distractor task. The distractor task consisted of a visual oddball paradigm that was run simultaneously with MMN, and the presentation of the visual stimuli were jittered to avoid co-occurring visual oddball and MMN ERP signals.

EEG Collection, Preprocessing, and ERP Averaging

Data Acquisition—EEG was recorded at 1024 Hz using either a 32-channel or 64-channel electrode cap. Additional electrodes were placed on the face and mastoids, and an offlineaverage mastoid reference was used for the following data analysis. In the present study, data from one subject on one test occasion were incomplete due to equipment malfunction, and required the elimination of some sections of the continuous recording

where shorting of the electrodes occurred. As a result, less than half of the total trials were available for this test session from the start of data preprocessing. For one additional subject and test occasion, operator error resulted in no continuous EEG recording in one of the three test blocks. Due to the complicated study design and small sample size, both of these recordings were included in all reliability analyses.

Preprocessing—EEG recordings were re-referenced to average mastoids and high-pass filtered at 1 Hz before being segmented into 1000 ms epochs (−500 to 500 ms). Blinks and eye movement artifacts were recorded by electrodes placed above and below the right eye (vertical electro-oculogram) and on the outer canthus of left and right eyes (horizontal electro-oculogram) were corrected for by using the ocular correction method outlined in Gratton, et al. (Gratton et al., 1983). Following baseline correction (−100 to 0 ms), outlier electrodes were interpolated within single trial epochs based on previously established criteria (Nolan et al., 2010). A spherical spline interpolation (Delorme and Makeig, 2004) was applied to any channel that was determined to be a statistical outlier ($|z| > 3$) on one or more of four parameters, including variance to detect additive noise, median gradient to detect high-frequency activity, amplitude range to detect pop-offs, and deviation of the mean amplitude from the common average to detect electrical drift. Epochs with amplitudes greater than ±100 μV in any of the following electrodes were rejected: AF3, AF4, F3, Fz, F4, FC1, FC2, FC5, FC6, C3, Cz, C4. Like previous studies (Biagianti et al., 2017, Fryer et al., 2020, Hay et al., 2015, Roach et al., 2020), these electrodes were selected because they either include or are spatially adjacent to the six, fronto-central electrodes of interest for the main NAPLS-2 MMN analyses.

ERP Averaging and MMN Measurement—ERP averages for all stimulus types were determined using a sorted averaging method (Rahne et al., 2008). This method has been shown to reduce noise in the MMN waveform by averaging over the subset of trials that optimizes the estimated signal to noise ratio (eSNR) for each subject. In this data set, single-epoch root mean squared (RMS) amplitude values at each of the 12 electrodes used for artifact rejection for each trial were calculated, averaged across electrode, and sorted in ascending order for each stimulus type. The subset of sorted trials selected for ERP averaging were associated with the largest eSNR, which is the ratio of the number of trials to the variance of the amplitude values across sorted trials. To facilitate comparison with more traditional MMN processing methods, a separate set of ERP averages were also obtained ommiting this sorted averaging step. Following averaging, ERPs for all stimulus types were low-pass filtered at 30 Hz, and then standard tone ERP waves were subtracted from deviants to obtain difference waves. As the reference electrode could influence both artifact rejection and sorted averaging trial elimination steps, nose re-referencing was done on the final waveforms to facilitate reference electrode comparisons on the exact same set of trials. MMN peak amplitude was classified as the most negative peak between 90 and 290 ms in each calculated difference wave. MMN mean amplitude ±10ms around the peak was also quantified as an alternative measurement to peak amplitude. Finally, average amplitude in a fixed windowbased on grand average waveforms (90-170ms for FRQ and DBL, 150-230 for DUR) was quantified as a third approach. Peak latencies were saved for a fourth set of generalizability analyses.

Traveling Subject Sample Reliability Analyses

Variance Components and G-coefficients—The main purpose of this fully crossed, two facet (site and test occasion) G-study design is to estimate variance components. Variance components can then be used to calculate generalizability or dependability (G- or D-) coefficients. The G-coefficient is relevant when relative measurements or differences between subjects are of interest (e.g., a 3 μV difference between MMN responses from a patient and control subject) while the D-coefficient is relevant when absolute measurements are of interest (e.g., a patient has a −3μV MMN response). The identification of critical facets and estimation of associated variance components is considered a G-study in the G-theory framework. While the G-study and estimated variance components are sufficient to calculate both G- and D-coefficients, the theory separately labels optimization of reliability coefficients for future studies or data collection procedures as a decision study (or D-study). In the D-study, estimated variance components are used to determine how many measurements are required to produce a sufficiently high G- or D-coefficient, when averaging across facets such as test item or occasion.

Variance is partitioned into the main random effects of Person, Site, and Occasion, their two-way interactions, and a final term corresponding to the three-way interaction plus error. This particular design allows one to estimate 7 variance components for any given score from each person, at each site, on each test occasion. The variance components and their definitions are described in Table 1.

Once variance components are estimated, the G-coefficient, which provides a measure of generalizability or reliability of the measured score, can be calculated as:

$$G = \frac{\sigma_p^2}{\left(\sigma_{\mathbf{p}}^2 + \frac{\sigma_{\mathbf{ps}}^2}{\mathbf{n}_s} + \frac{\sigma_{\mathbf{po}}^2}{\mathbf{n_0}} + \frac{\sigma_{\mathbf{pso}+\mathbf{e}}^2}{\mathbf{n_s n_o}}\right)}$$

The larger NAPLS-2 parent study design included EEG assessments at baseline, 12 month, and 24 month study time points. Since MMN scores from each session would be treated separately, with particular emphasis on using baseline data to predict conversion to psychosis in the parent study, the logical choice for $\boldsymbol{n_o}$ is 1. Likewise, since all subjects would only be studied at their home site, the logical choice for $\boldsymbol{n_s}$ is 1. Therefore, the G-coeffecient is equivalent to the intraclass correlation (ICC) as defined by Shrout and Fleiss (e.g., ICC(3,1) in (Shrout and Fleiss, 1979)) when $\boldsymbol{n_o}$=$\boldsymbol{n_s}$=1. Variance components were estimated using a restricted maximum likelihood approach implemented in Matlab (Witkovský, 2012). Components were estimated and saved separately for each deviant type (DBL, FRQ, DUR), electrode (32 from overlapping montage), MMN measurement (peak amplitude, mean around peak, mean in fixed window, and peak latency), ERP averaging method (sorted or traditional) and reference electrode (average mastoids or nose). Tables in the main text focus on G-coefficients averaged across the six, fronto-central electrodes (F3, Fz, F4, C3, Cz, C4) of interest for the main NAPLS-2 MMN analyses.

While the visual oddball attention task and hearing tests are not the focus of this report, variance components were also estimated for median target reaction time (RT), overall accuracy in the oddball task, and mean hearing thresholds from the hearing test.

Additional Generalizability Analyses—As the most frequently reported MMN reliability studies collect data on two test occassions at one laboratory site, 9 separate sets of reduced variance components were estimated for a single facet (test occasion) crossed design within each of the eight sites, and a final model where each subject's initial pair of test occasions from their home site were used ("home" site model). In this last model, person and site are completely confounded, so associated variance components and G coefficients must be interpreted with this limitation in mind.

Results

Variance Components and G-coefficients

MMN ERP waveforms from electrode Fz are plotted in Figure 1. There is clear similarity between waveforms at each site and on each test occasion up until about 200 ms, followed by greater variability in the 200-400 ms range.

A similar set of waveforms was produced for the traditional averaging approach (Supplementary Figure S1). Accepted trial numbers for the two averaging approaches are included in Table 2. Sorted averaging resulted in the rejection of between 6% - 7% of the trials for each trial type on average. In some sessions, sorted averaging rejected no additional trials for the deviants (range: 0 – 18 trials), while at least 10 trials were rejected from the standards in each session (range: 10 – 218 trials).

Figure 2 shows the grand average MMN waveforms from electrode Fz across all 128 sessions along with G-coefficient waveforms. While the sorted and traditional grand average ERPs are almost identical, the G-coefficient waveforms are less consistent between methods with all samples falling well below 0.4, indicating poor reliability. The peak and mean around the peak G-coefficients for Fz are greater (0.275-0.487) than any G-coefficient in the waveforms, indicating that latency jitter in the MMN may contribute to poor sample-wise reliability in the waveforms.

Table 3 lists the proportion of variance for each of the 7 variance components as well as G-coefficients averaged across the 6 fronto-central electrodes (F3, Fz, F4, C3, Cz, C4) and deviants. Table 4 presents these averaged G-coefficients, separated by deviant type. Such averaging is consistent with the group analysis approach applied to MMN data in other CHR(Bodatsch et al., 2011, Perez et al., 2014) and schizophrenia(Doring et al., 2016, Duncan et al., 2009, Hamilton et al., 2018, Hay et al., 2015) studies. G-coefficients for each electrode, deviant type, reference, averaging approach, and measure are included in Supplementary Table S1. They were less than what was expected (i.e., almost all G-coefficients < 0.6).

As shown in Tables 3 and 4, the two averaging approaches yielded similar estimates. This is consistent with data from Fz only as plotted in Figure 3, which shows similar percentages of

variance attributed to each of the 7 variance components, when additionally separated by each deviant type. The general pattern shows that the largest variance component is error, followed by Person and then Person X Site for the three amplitude measures. For the peak latency measure, error variance still dominates, but Person X Site variance is greater than person variance.

Topographic maps on the mean amplitude in the fixed time windows and their corresponding G-coefficients for mastoid referenced, sorted average data are shown in Figure 4. Corresponding plots for traditional average data and peak amplitude are included as Supplementary Figures S2–S4. Much like the data from electrode Fz, the Window measure reliability follows a pattern of DUR > DBL > FRQ. All three deviant types exhibit a left-lateralized central-parietal reliability maximum, and the topographies of the scored amplitude do not perfectly match the associated G-coefficient topographies.

Overall oddball accuracy reliability was poor (G = 0.2327), given the overall lack in variability and ceiling level performance across most sessions (Median Accuracy = 100%, inter-quartile range: 99.77% - 100%). Median target RT (Median RT = 364.7, inter-quartile range: 353.3 to 393.5ms) reliability was good (G = 0.6505); the only non-zero variance components in addition to Person and the Error terms were Site and Person X Site (2.45 and 11.92% variance explained, respectively). Taken together, these measures indicate that subjects performed the visual distraction task consistently across sessions. The mean hearing level reliability was fair (G = 0.4961) with a larger proportion of variance attributed to Site (12.01%) than most other measures studied. Person X Site and Site X Occasion variance components were also non-zero (7.77 and 2.08% variance explained, respectively).

Additional Site-Specific Generalizability Analyses

A similar pattern can be seen when the G-coefficients are calculated separately on a per site basis (see Figure 5) using a reduced, single-facet (Occasion) model. In Table 5, G-coefficients calculated separately within each Site and averaged across the fronto-central 6 electrodes of interest in NAPLS-2 analyses are presented side by side for sorted and traditional averaging approaches for the Window measure.

While these site-specific G-coefficients were greater than those observed for the two-facet models, they were less than what was expected. G-coefficients for each electrode, deviant type, reference, averaging approach, and measure calculated separately at each of the 8 sites (and a 9^{th} set for “home” site) are included in Supplementary Table S2.

The G-coefficients for both sorted and traditional averaging approaches were comparable, and the Window measure G-coefficients were similar to both Peak and Mean measures of MMN amplitude, indicating that the planned analytic approach of using the sorted averaging Window measure to quantify and assess MMN in CHR and comparison controls is appropriate. However, nose referencing the data resulted in reduction in G-coefficients in approximately 75% of electrode, measure, and averaging combinations, and the majority (~94%) of G-coefficients from nose-referenced data were poor (i.e., Gs < .04, see Supplementary Tables S1 and S2).

Discussion

The main purpose of this generalizability study was to quantify variance components and associated G-coefficients representing the single site, single session reliability of MMN quantified with the same approach planned for the larger data set and clinical comparisons in NAPLS-2. An additional goal was to compare this planned MMN scoring approach to alternative methods of referencing, averaging, and scoring MMN data. Across these different two-facet models, error variance was typically the largest, followed by either person or person by site variance components, depending on the type of response being measured (i.e., amplitude vs latency). All other variance components accounted for small proportions of variance in the study data. While there were no major differences between averaging approaches or MMN amplitude scoring methods, nose referencing had a negative impact on reliability. The majority of G-coefficients were less than expected given test-retest reliability reported in prior single site studies. Finally, G-coefficients were calculated separately for each of the 8 NAPLS-2 sites and a “home” site model to allow for qualitative comparisons within the consortium. These single laboratory site G-coefficients were comparable to prior studies, but reliability was less consistent across NAPLS-2 sites than expected.

Comparisons of MMN reliability using average mastoid versus nose references revealed that mastoid referenced data were more reliable in the majority (~75%) of electrodes, deviants, measurements, and averaging approaches. Many test-retest reliability studies of MMN have used nose referenced data, most likely to showthat the MMN component reverses polarity and that the associated scalp component is the MMN and not the N2b ERP component. Based on the findings from the current study, nose referencing appears to quite clearly increase relative error variance. Therefore, reports that MMN suffers from low or poor test-retest reliability that were based on a nose reference should be qualified as limited to MMN measures calculated using this particular reference. However, it should also be noted that prior MMN reliability studies with nose referencing predominantly used low-impedance recordings, which could improve nose-referenced data quality.

The comparisons of reliability of MMN scores using two different averaging methods had mixed results. The sorted averaging approach, which removed an additional 6% of trials on average, only improved reliability estimates compared to a traditional averaging approach in slightly more than half (~55%) of electrodes, deviants, and measurements. However, when limiting the focus to fronto-central electrodes and amplitude measures typically used in MMN group analyses, the difference in percentage of variance attributed to persons was less than 1.5%, on average, for these two averaging approaches. It is posible that the benefit of the sorted averaging algorithm is limited to electrodes where the signal is smaller, but this benefit seems to be very small, especially when one considers that the computation time and single trial implementation of the sorted averaging algorithm may be prohibitive for many EEG researchers.

Two previous studies reported excellent reliability (Fz ICCs> 0.85) using a long duration deviant similar to that used in the present study based on a window measurement (135-205ms) from nose-referenced data (Light and Braff, 2005, Light et al., 2012). Thesehigh reliability coefficients were based on either 10 patients with schizophrenia tested

twice, 18 months apart (Light and Braff, 2005) or 163 patients and 58 comparison controls tested after 1 year (Light et al., 2012). Notably, the studies by Light et al used a much longer recording session that was terminated after a minimum of 225 artifact free deviant trials were obtained for each subject at run time and usually resulting in >250 trials following post-acquisition artifact correction procedures. While the corresponding traditionally averaged, mastoid referenced, window measure G-coefficient was smaller in the present study (Fz G = 0.492), the reduced "home" site model was almost identical (Fz G_{home} = 0.883). Another previous long duration deviant MMN reliability study of 19 healthy subjects tested twice, 7 to 56 days apart, had good reliability (Fz ICC = 0.66) using a left earlobe reference and similar window (50-200ms) measurement (Hall et al., 2006). Lew et al. (Lew et al., 2007) also reported good reliability (Cz ICC = 0.6) in data from 19 healthy subjects tested twice, 2 to 60 days apart, using a frequency deviant and nose reference. However, unlike the previous two studies and the current G-study, the tones were part of an active auditory oddball attention task.

There are several limitations to the current study that should be carefully considered. First, estimates of variance components can be fairly unstable when the number of observations is small, and the estimates may have been impacted by having only 8 subjects studied on only two test occassions at each site. While our exact design is unlikely to be replicated in future reliability studies, focusing on a larger sample size and more repeat test occassions would yield more stable variance component estimates and potentially more informative decision studies. In addition to this sample size being small, the nature of this traveling study design required recruiting adult subjects who were able to complete 16 days of testing and travel across North America. One indicator that this sample may not be representative of the larger NAPLS-2 sample, or yet another indicator that this sample size was too small, is that the variance in MMN responses in this traveling subject sample was reduced relative to the variance in our larger (N=241) healthy control sample used in the parent NAPLS-2 MMN analyses (Roach et al., 2020). Specifically, FRQ MMN variance was 3.32μV in the larger group and 1.43μV in this travel sample; DBL MMN variance was 4.29μV in the larger group and 2.02μV in this travel sample; and DUR MMN variance was 4.02μV in the larger group and 2.29μV in this travel sample. The limitations of this study due to small sample size cannot be overstated.

Second, the particular design employed here could have also introduced unintended psychological and/or physiological effects on the ERP measures. The participants completed the same EEG task 16 times, and 14 of these 16 test occasions involved some long travel times, which could have contributed to boredom, sleepiness, jetlag, and/or stress. This constrasts with previous MMN reliability studies that involve typically two, but at most four (Dalebout and Fox, 2001) or five (Paukkunen et al., 2011), repeated assessments at one lab site. The geographic layout of the sites and administrative burden of organizing the study required a fixed travel loop for all subjects, and pseudo-randomization of order was achieved by having one subject start at each site. For example, all subjects visited UCSD after UCLA except for the subject who started at UCSD. While order effects were not anticipated, they cannot be quantified in the current design and may have contributed to the variability of the within site reliability coefficients and the large Person X Site variance components in the fully crossed, two facet models. Based on suggestions from one reviewer, additional models

were run including time of day, time since last EEG session, and vigilance factors, but no additional factor explained a significant proportion of MMN variability (see Supplementary Material). Despite these limitations, MMN measures have equal or greater reliability than task-based fMRI measures from this same cohort (Forsyth et al., 2014, Gee et al., 2015).

In conclusion, the current study demonstrates the feasibility of a multisite, EEG studies of mismatch negativity using the same software and hardware. Grand average ERP waveforms were consistent across all NAPLS sites and test occasions (Figure 1) despite the small sample size and high number of repeated assessments conducted on each subject over a relatively brief assessment period. Moreover, when only the first two test occasions were considered from each traveling subject's "home" site, reliability was good or better (i.e., G-coefficient or ICC > 0.6) in a large proportion of MMN measurements across electrodes. This home site design closely matches the main NAPLS design, in which subjects participate at a single site. Investigators should seriously consider the value of future traveling subject EEG studies given the relatively high cost and low sample size. Given the highly variable but generally poor reliability of latency measures, MMN amplitude assessed using a fixed latency, mean amplitude as planned for NAPLS-2 MMN clinical analyses (e.g., "Window" measure in this study) may be the most generalizable measure for multisite investigations of MMN in low prevalance patient groups, such as CHR individuals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by grants from National Institute of Mental Health (U01MH081902 to TDC, P50 MH066286 to CEB, U01MH081988 to EFW, U01MH076989to DHM, U01MH081944to KSC, U01MH081984 to JA, U01MH082004 to DOP, U01MH081857 to BAC, U01MH081928 to LJS, U01MH082022 to SW).

Financial Disclosures

Dr Light reported grants from Boehringer Ingelheim, other from Astellas, and other from Heptares outside the submitted work. Dr Bearden reported grants from the NIMH during the conduct of the study. Dr Cornblatt reported grants from NIMH during the conduct of the study. Dr Duncan has received research support for work unrelated to this project from Auspex Pharmaceuticals, Inc. and Teva Pharmaceuticals, Inc. Dr Perkins reported grants from the NIMH during the conduct of the study; personal fees from Sunovion and personal fees from Alkermes outside the submitted work. Dr Seidman reported grants from the NIMH during the conduct of the study. Dr Woods reported grants from the NIMH during the conduct of the study; grants and personal fees from Boehringer Ingelheim, personal fees from New England Research Institute, personal fees from Takeda, grants from Amarex, grants from Teva, grants from One Mind Institute, and grants from Substance Abuse and Mental Health Services Administration outside the submitted work; in addition, Dr Woods had a patent to Glycine agonists for prodromal schizophrenia issued and a patent to Method of predicting psychosis risk using blood biomarker analysis pending. Dr Cannon reported grants from NIMH during the conduct of the study. Dr Mathalon reported grants from NIMH during the conduct of the study; consulting fees from Boehringer Ingelheim, consulting fees from Aptinyx, consulting fees from Takeda, consulting fees from Upsher-Smith, and consulting fees from Alkermes outside the submitted work. No other disclosures were reported.

References

Addington J, Cadenhead KS, Cornblatt BA, Mathalon DH, McGlashan TH, Perkins DO, et al. North American Prodrome Longitudinal Study (NAPLS 2): overview and recruitment. Schizophr Res 2012;142(1-3):77–82. [PubMed: 23043872]

Andre-Obadia N, Zyss J, Gavaret M, Lefaucheur JP, Azabou E, Boulogne S, et al. Recommendations for the use of electroencephalography and evoked potentials in comatose patients. Neurophysiol Clin 2018;48(3):143–69. [PubMed: 29784540]

Avissar M, Xie S, Vail B, Lopez-Calderon J, Wang Y, Javitt DC. Meta-analysis of mismatch negativity to simple versus complex deviants in schizophrenia. Schizophr Res 2018;191:25–34. [PubMed: 28709770]

Biagianti B, Roach BJ, Fisher M, Loewy R, Ford JM, Vinogradov S, et al. Trait aspects of auditory mismatch negativity predict response to auditory training in individuals with early illness schizophrenia. Neuropsychiatr Electrophysiol 2017;3:2. [PubMed: 28845238]

Bodatsch M, Ruhrmann S, Wagner M, Muller R, Schultze-Lutter F, Frommann I, et al. Prediction of psychosis by mismatch negativity. Biol Psychiatry 2011;69(10):959–66. [PubMed: 21167475]

Brown GG, Mathalon DH, Stern H, Ford J, Mueller B, Greve DN, et al. Multisite reliability of cognitive BOLD data. Neuroimage 2011;54(3):2163–75. [PubMed: 20932915]

Cadenhead KS, Addington J, Cannon TD, Cornblatt BA, de la Fuente-Sandoval C, Mathalon DH, et al. Between-site reliability of startle prepulse inhibition across two early psychosis consortia. Neuroreport 2013;24(11):626–30. [PubMed: 23799460]

Cannon TD, Sun F, McEwen SJ, Papademetris X, He G, van Erp TG, et al. Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis. Hum Brain Mapp 2014;35(5):2424–34. [PubMed: 23982962]

Chen C, Chan CW, Cheng Y. Test-Retest Reliability of Mismatch Negativity (MMN) to Emotional Voices. Front Hum Neurosci 2018;12:453. [PubMed: 30498437]

Dalebout SD, Fox LG. Reliability of the mismatch negativity in the responses of individual listeners. J Am Acad Audiol 2001;12(5):245–53. [PubMed: 11392436]

Danjou P, Viardot G, Maurice D, Garces P, Wams EJ, Phillips KG, et al. Electrophysiological assessment methodology of sensory processing dysfunction in schizophrenia and dementia of the Alzheimer type. Neurosci Biobehav Rev 2019;97:70–84. [PubMed: 30195932]

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004;134(1):9–21. [PubMed: 15102499]

Deouell LY, Bentin S. Variable cerebral responses to equally distinct deviance in four auditory dimensions: a mismatch negativity study. Psychophysiology 1998;35(6):745–54. [PubMed: 9844436]

Doring C, Muller M, Hagenmuller F, Ajdacic-Gross V, Haker H, Kawohl W, et al. Mismatch negativity: Alterations in adults from the general population who report subclinical psychotic symptoms. Eur Psychiatry 2016;34:9–16. [PubMed: 26928341]

Duncan CC, Barry RJ, Connolly JF, Fischer C, Michie PT, Naatanen R, et al. Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. Clin Neurophysiol 2009;120(11):1883–908. [PubMed: 19796989]

Erickson MA, Ruffle A, Gold JM. A Meta-Analysis of Mismatch Negativity in Schizophrenia: From Clinical Risk to Disease Specificity and Progression. Biol Psychiatry 2016;79(12):980–7. [PubMed: 26444073]

First MB, Spitzer RL, Gibbon M & Williams JBW Structured Clinical Interview for DSM-IV Axis I Disorders (SCID), Research Version, Patient Edition with Psychotic Screen. New York: Biometrics Research, New York State Psychiatric Institute, 1997.

Forsyth JK, McEwen SC, Gee DG, Bearden CE, Addington J, Goodyear B, et al. Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: analysis from the North American Prodrome Longitudinal Study. Neuroimage 2014;97:41–52. [PubMed: 24736173]

Fryer SL, Roach BJ, Hamilton HK, Bachman P, Belger A, Carrion RE, et al. Deficits in auditory predictive coding in individuals with the psychosis risk syndrome: Prediction of conversion to psychosis. J Abnorm Psychol 2020;129(6):599–611. [PubMed: 32757603]

Gee DG, McEwen SC, Forsyth JK, Haut KM, Bearden CE, Addington J, et al. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study. Hum Brain Mapp 2015;36(7):2558–79. [PubMed: 25821147]

Gratton G, Coles MG, Donchin E. A new method for off-line removal of ocular artifact. Electroencephalogr Clin Neurophysiol 1983;55(4):468–84. [PubMed: 6187540]

Hall MH, Schulze K, Rijsdijk F, Picchioni M, Ettinger U, Bramon E, et al. Heritability and reliability of P300, P50 and duration mismatch negativity. Behav Genet 2006;36(6):845–57. [PubMed: 16826459]

Hamilton HK, Perez VB, Ford JM, Roach BJ, Jaeger J, Mathalon DH. Mismatch Negativity But Not P300 Is Associated With Functional Disability in Schizophrenia. Schizophr Bull 2018;44(3):492–504. [PubMed: 29036701]

Hay RA, Roach BJ, Srihari VH, Woods SW, Ford JM, Mathalon DH. Equivalent mismatch negativity deficits across deviant types in early illness schizophrenia-spectrum patients. Biol Psychol 2015;105:130–7. [PubMed: 25603283]

Hermens DF, Chitty KM, Kaur M. Mismatch negativity in bipolar disorder: A neurophysiological biomarker of intermediate effect? Schizophr Res 2018;191:132–9. [PubMed: 28450056]

Horvath A, Szucs A, Csukly G, Sakovics A, Stefanics G, Kamondi A. EEG and ERP biomarkers of Alzheimer's disease: a critical review. Front Biosci (Landmark Ed) 2018;23:183–220. [PubMed: 28930543]

Kathmann N, Frodl-Bauch T, Hegerl U. Stability of the mismatch negativity under different stimulus and attention conditions. Clin Neurophysiol 1999;110(2):317–23. [PubMed: 10210621]

Klem GH, Luders HO, Jasper HH, Elger C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 1999;52:3–6. [PubMed: 10590970]

Kujala T, Kallio J, Tervaniemi M, Naatanen R. The mismatch negativity as an index of temporal processing in audition. Clin Neurophysiol 2001;112(9):1712–9. [PubMed: 11514254]

Kujala T, Leminen M. Low-level neural auditory discrimination dysfunctions in specific language impairment-A review on mismatch negativity findings. Dev Cogn Neurosci 2017;28:65–75. [PubMed: 29182947]

Lew HL, Gray M, Poole JH. Temporal stability of auditory event-related potentials in healthy individuals and patients with traumatic brain injury. J Clin Neurophysiol 2007;24(5):392–7. [PubMed: 17912063]

Light GA, Braff DL. Stability of mismatch negativity deficits and their relationship to functional impairments in chronic schizophrenia. Am J Psychiatry 2005;162(9):1741–3. [PubMed: 16135637]

Light GA, Swerdlow NR, Rissling AJ, Radant A, Sugar CA, Sprock J, et al. Characterization of neurophysiologic and neurocognitive biomarkers for use in genomic and clinical outcome studies of schizophrenia. PLoS One 2012;7(7):e39434. [PubMed: 22802938]

McCleery A, Mathalon DH, Wynn JK, Roach BJ, Hellemann GS, Marder SR, et al. Parsing components of auditory predictive coding in schizophrenia using a roving standard mismatch negativity paradigm. Psychol Med 2019:1–12.

McGlashan T, Walsh B, Woods S. The psychosis-risk syndrome: handbook for diagnosis and follow-up: Oxford University Press, 2010.

Miller TJ, McGlashan TH, Rosen JL, Somjee L, Markovich PJ, Stein K, et al. Prospective diagnosis of the initial prodrome for schizophrenia based on the Structured Interview for Prodromal Syndromes: preliminary evidence of interrater reliability and predictive validity. Am J Psychiatry 2002;159(5):863–5. [PubMed: 11986145]

Naatanen R, Shiga T, Asano S, Yabe H. Mismatch negativity (MMN) deficiency: a break-through biomarker in predicting psychosis onset. Int J Psychophysiol 2015;95(3):338–44. [PubMed: 25562834]

Noble S, Scheinost D, Finn ES, Shen X, Papademetris X, McEwen SC, et al. Multisite reliability of MR-based functional connectivity. Neuroimage 2017;146:959–70. [PubMed: 27746386]

Nolan H, Whelan R, Reilly RB. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. J Neurosci Methods 2010;192(1):152–62. [PubMed: 20654646]

Paukkunen AK, Leminen M, Sepponen R. The effect of measurement error on the test-retest reliability of repeated mismatch negativity measurements. Clin Neurophysiol 2011;122(11):2195–202. [PubMed: 21570906]

Pekkonen E, Rinne T, Naatanen R. Variability and replicability of the mismatch negativity. Electroencephalogr Clin Neurophysiol 1995;96(6):546–54. [PubMed: 7489676]

Perez VB, Woods SW, Roach BJ, Ford JM, McGlashan TH, Srihari VH, et al. Automatic auditory processing deficits in schizophrenia and clinical high-risk patients: forecasting psychosis risk with mismatch negativity. Biol Psychiatry 2014;75(6):459–69. [PubMed: 24050720]

Rahne T, von Specht H, Muhler R. Sorted averaging--application to auditory event-related responses. J Neurosci Methods 2008;172(1):74–8. [PubMed: 18499265]

Ramlakhan JU, Zomorrodi R, Downar J, Blumberger DM, Daskalakis ZJ, George TP, et al. Using Mismatch Negativity to Investigate the Pathophysiology of Substance Use Disorders and Comorbid Psychosis. Clin EEG Neurosci 2018;49(4):226–37. [PubMed: 29502434]

Recasens M, Uhlhaas PJ. Test-retest reliability of the magnetic mismatch negativity response to sound duration and omission deviants. Neuroimage 2017;157:184–95. [PubMed: 28576412]

Roach BJ, Hamilton HK, Bachman P, Belger A, Carrion RE, Duncan E, et al. Stability of mismatch negativity event-related potentials in a multisite study. Int J Methods Psychiatr Res 2020;29(2):e1819. [PubMed: 32232944]

Schall U, Catts SV, Karayanidis F, Ward PB. Auditory event-related potential indices of fronto-temporal information processing in schizophrenia syndromes: valid outcome prediction of clozapine therapy in a three-year follow-up. Int J Neuropsychopharmacol 1999;2(2):83–93. [PubMed: 11281974]

Schroger E, Giard MH, Wolff C. Auditory distraction: event-related potential and behavioral indices. Clin Neurophysiol 2000;111(8):1450–60. [PubMed: 10904227]

Schwartz S, Shinn-Cunningham B, Tager-Flusberg H. Meta-analysis and systematic review of the literature characterizing auditory mismatch negativity in individuals with autism. Neurosci Biobehav Rev 2018;87:106–17. [PubMed: 29408312]

Shaikh M, Valmaggia L, Broome MR, Dutt A, Lappin J, Day F, et al. Reduced mismatch negativity predates the onset of psychosis. Schizophr Res 2012;134(1):42–8. [PubMed: 22024244]

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8. [PubMed: 18839484]

Tervaniemi M, Lehtokoski A, Sinkkonen J, Virtanen J, Ilmoniemi RJ, Naatanen R. Test-retest reliability of mismatch negativity for duration, frequency and intensity changes. Clin Neurophysiol 1999;110(8):1388–93. [PubMed: 10454274]

Umbricht D, Krljes S. Mismatch negativity in schizophrenia: a meta-analysis. Schizophr Res 2005;76(1):1–23. [PubMed: 15927795]

Uwer R, von Suchodoletz W. Stability of mismatch negativities in children. Clin Neurophysiol 2000;111(1):45–52. [PubMed: 10656510]

Volkmer S, Schulte-Korne G. Cortical responses to tone and phoneme mismatch as a predictor of dyslexia? A systematic review. Schizophr Res 2018;191:148–60. [PubMed: 28712970]

Witkovsky V Estimation, Testing, and Prediction Regions of the Fixed and Random Effects by Solving the Henderson's Mixed Model Equations. Meas Sci Rev 2012;12(6):234–248.

Highlights

- The Mismatch Negativity (MMN) shows fair test, re-test reliability across 8 geographic testing sites.
- MMN reliability is greater using average mastoid than nose as the reference electrode.
- Multisite EEG studies of rare patient groups using MMN are feasible.

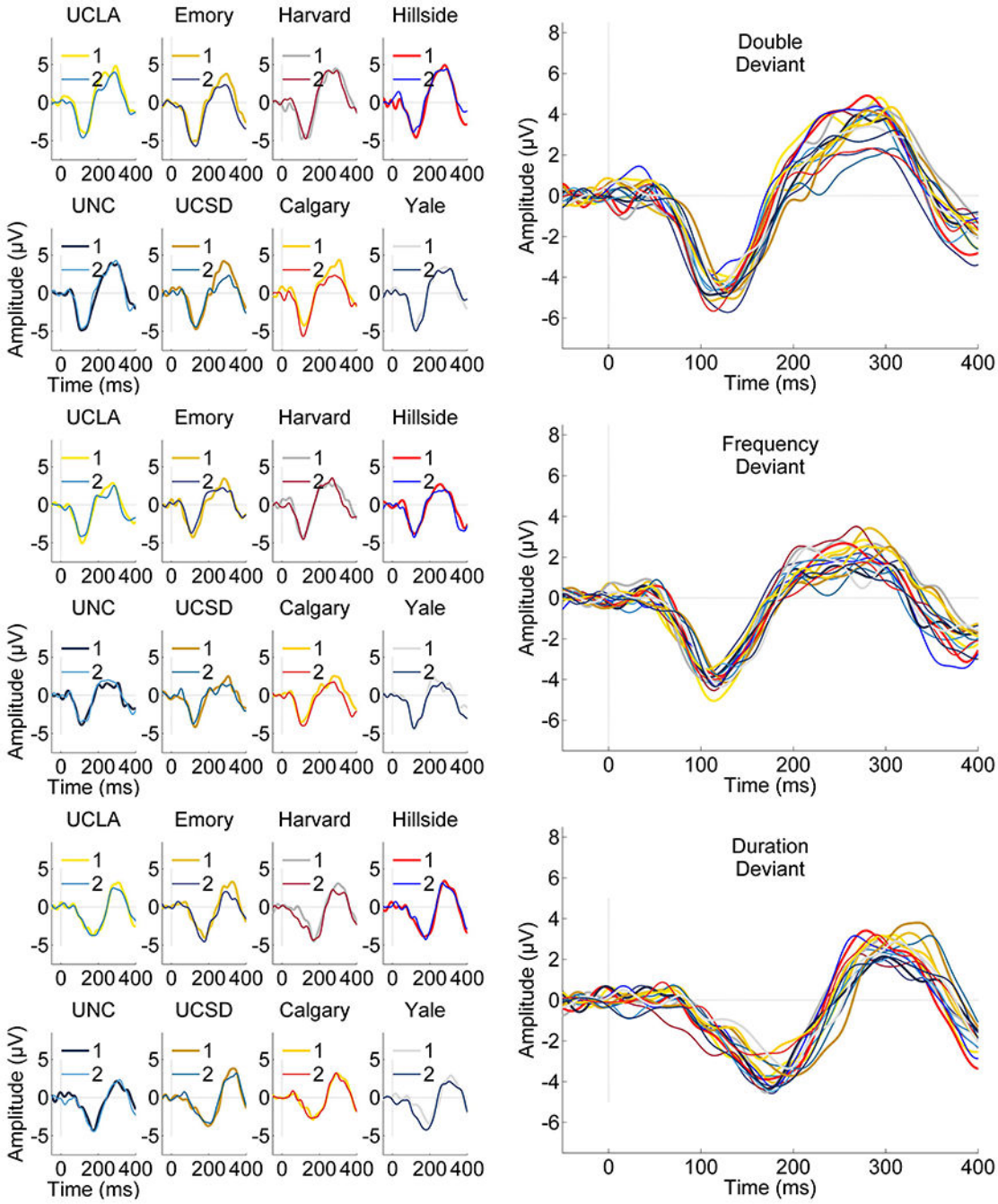

Figure 1.

Site and session-specific grand average mismatch negativity (MMN) deviant minus standard tone difference waveforms are plotted for the Double (Frequency plus Duration) Deviant (Top), Frequency Deviant (Middle), and Duration Deviant (Bottom) from electrode Fz. Grand Average MMN waveforms for each NAPLS laboratory site are plotted separately on the right-hand side for the first (1) and second (2) test occasion. All 16 of these average waveforms are overlaid for each deviant type on the left-hand side. Time, in milliseconds

(ms) from tone onset is plotted on the x-axis, and amplitude, in microVolts (μV), is plotted on the y-axis. Individual averages were 30Hz low-pass filtered prior to grand averaging.

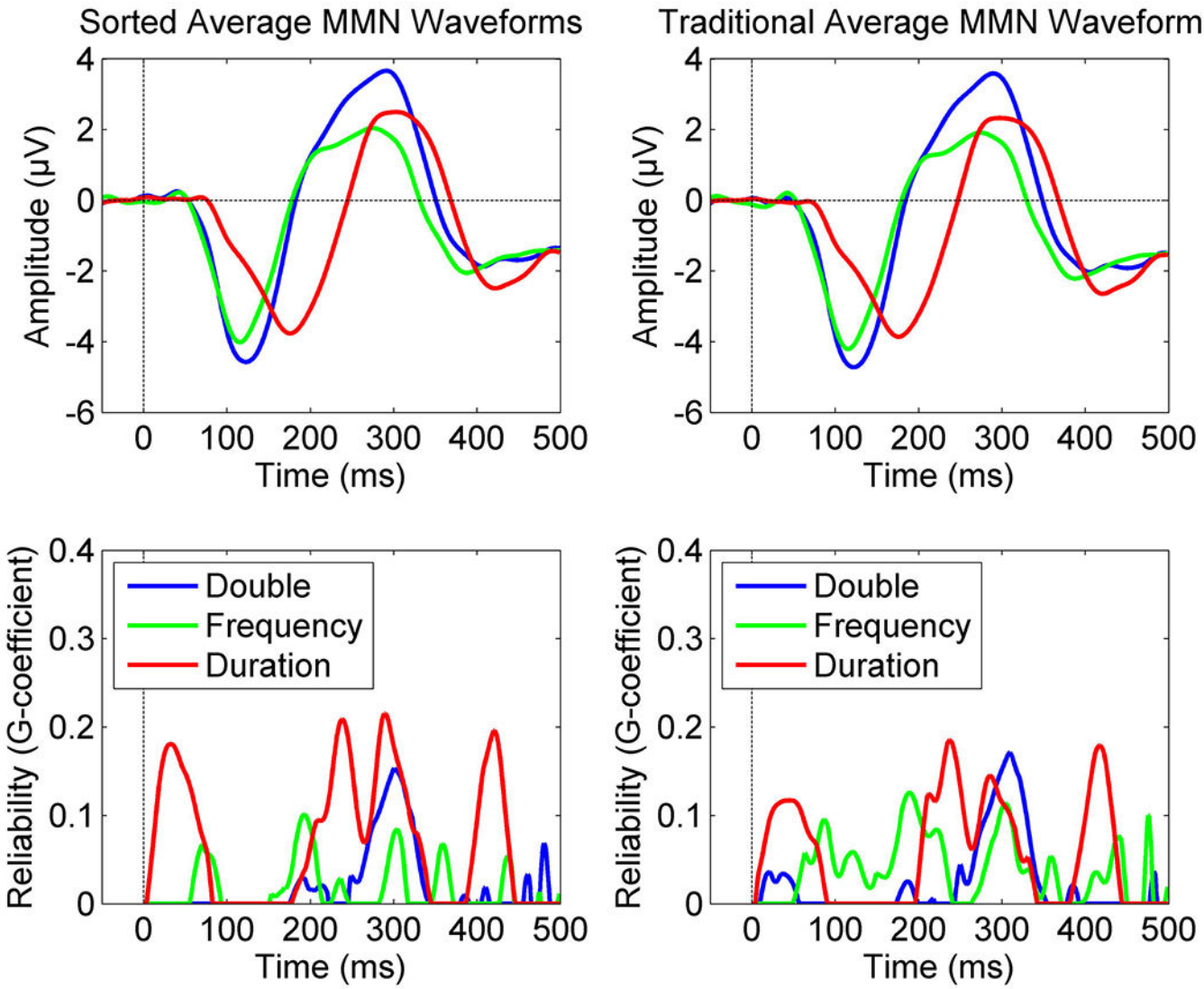

Figure 2.

In the top row, grand average mismatch negativity (MMN) deviant minus standard tone difference waveforms are plotted for the Double (Frequency plus Duration) Deviant (blue), Frequency Deviant (green), and Duration Deviant (red) from electrode Fz. Sorted averaging (left) and traditional averaging (right) MMN waveforms are very similar in these grand averages across all 128 test sessions. Individual averages were 30Hz low-pass filtered prior to grand averaging. Test-retest reliability waveforms are plotted separately for each deviant type and averaging method in the bottom row. These g-coefficient waveforms are derived from a two-facet (site and test occasion) fully-crossed generalizability analysis and demonstrate that the reliability (or generalizability) of the MMN waveform at any given time sample is relatively poor, indicating that MMN scores should be calculated with some averaging across time samples or peak-picking approach.

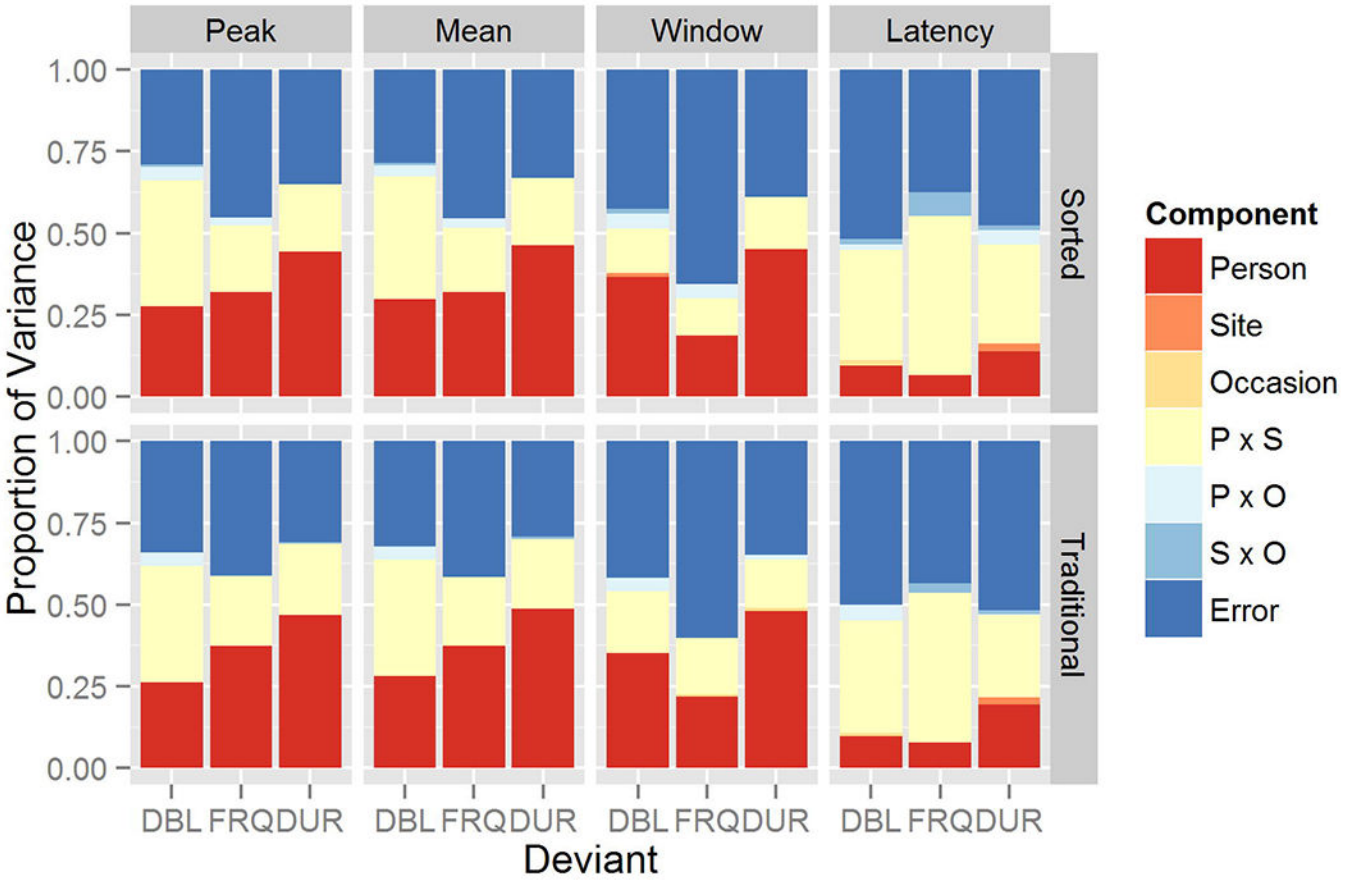

Figure 3.

Stacked bar plots show the proportion of variance (y-axis) explained by the 7 different variance components estimated using the two-facet, fully-crossed models. The Person (red), Person x Site (yellow), and Residual Error (dark blue) variance components account for the most of the variance for mismatch negativity peak amplitude (far left), mean amplitude ± 10 milliseconds around the peak (middle left), mean amplitude in a fixed time window (middle right), and peak latency (far right) measures. Double (Frequency plus Duration; DBL), Frequency (FRQ), and Duration (DUR) Deviants are plotted separately along the x-axis from electrode Fz, and are separated by sorted averaging (top row) and traditional averaging (bottom row) methods of event-related potential calculation.

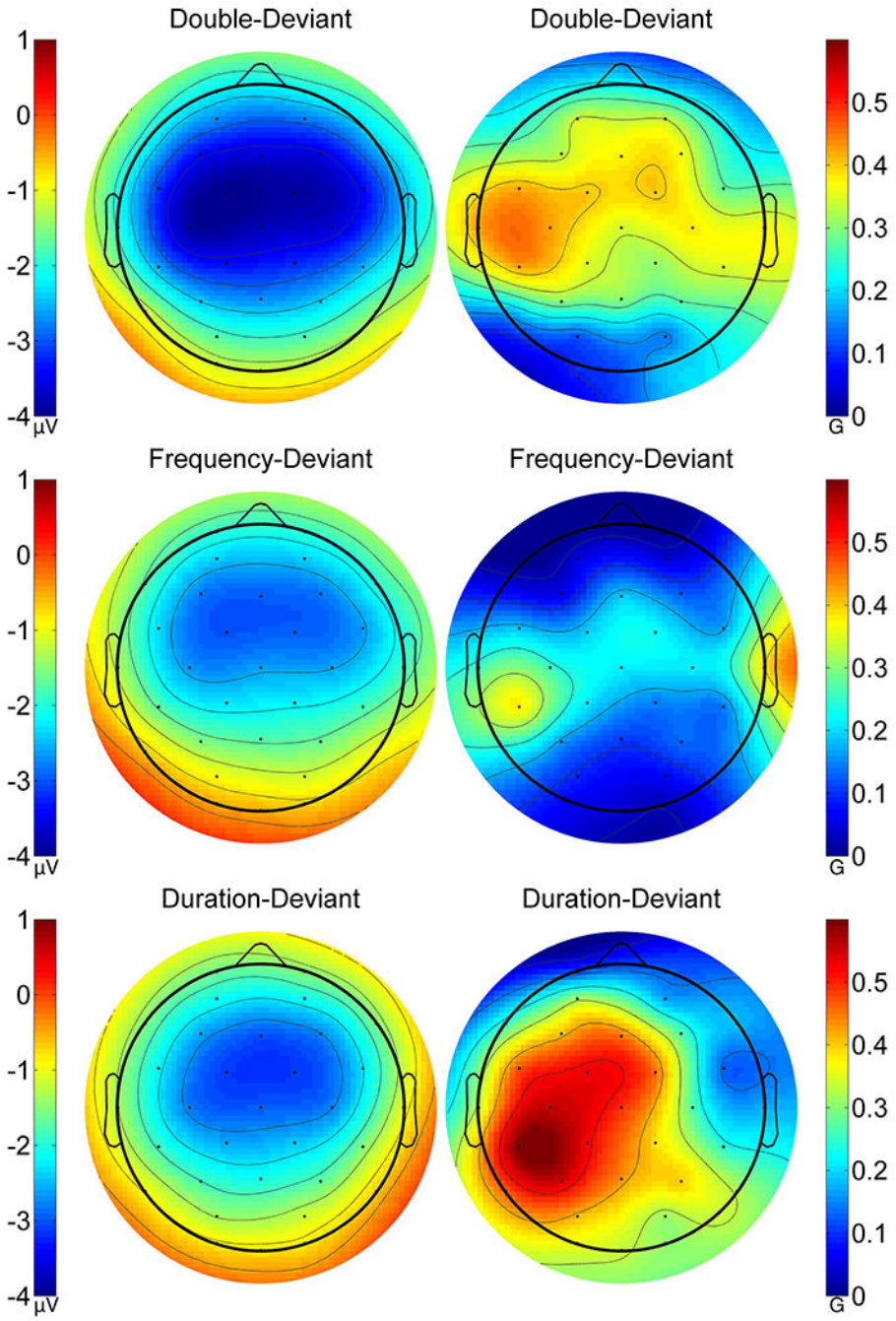

Figure 4.

Scalp topographic maps displaying the mismatch negativity (MMN) quantified as the mean amplitude (in microVolts, μV) in fixed latency windows on the left-hand side, and corresponding G-coefficients for the fully crossed two-facet (site and test occasion) generalizability study on the right-hand side. Frequency deviant (middle row) and double (duration+frequency) deviant MMN is averaged across 90-170 milliseconds (ms) while Duration deviant (bottom row) MMN is averaged across 150-230 ms.

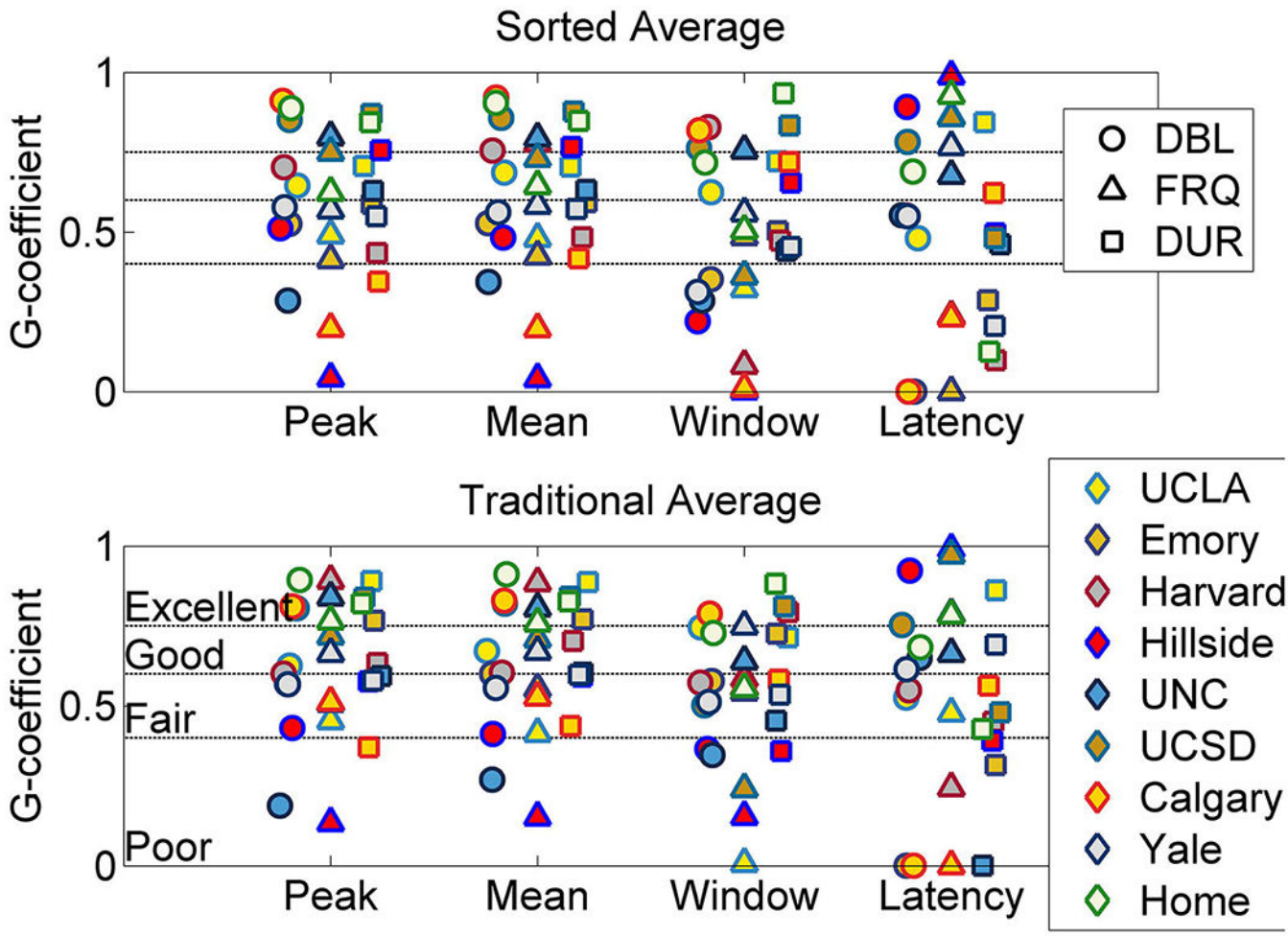

Figure 5.

G-coefficients for the single-facet (test occasion) generalizability sub-studies calculated separately for each NAPLS geographic site and a 9th "home" site G-study for electrode Fz based on either sorted averaging (top) or traditional averaging (bottom) event-related potential calculation. In all cases, measurement approaches are plotted along the x-axis separately for double-deviant (DBL, circles), frequency-deviant (FRQ, triangles), and duration-deviant (DUR, squares) mismatch negativity.

Table 1.

Variance Components for Person(p) x Site(s) x Occasion(o) Fully Crossed Design

Random Effect	V.C.	Expected Mean Squares (EMS) Definition
Person	σ_p^2	$\frac{EMS(p) - EMS(ps) - EMS(po) + EMS(pso)}{n_s n_o}$
Site	σ_s^2	$\frac{EMS(s) - EMS(ps) - EMS(so) + EMS(pso)}{n_p n_o}$
Occasion	σ_o^2	$\frac{EMS(s) - EMS(po) - EMS(so) + EMS(pso)}{n_p n_s}$
Person x Site	σ_{ps}^2	$\frac{EMS(ps) - EMS(pso)}{n_o}$
Person x Occasion	σ_{po}^2	$\frac{EMS(po) - EMS(pso)}{n_s}$
Site x Occasion	σ_{so}^2	$\frac{EMS(so) - EMS(pso)}{n_p}$
Person x Site x Occasion + Error	$\sigma_{pso}^2 + e$	$EMS(pso + e)$

Table 2.

Trial Numbers

Trial Type	Traditional Averaging Trial Numbers	Percentage removed by artifact rejection	Sorted Averaging Trial Numbers	Percentage removed by sorting
Standard	1468.13 ± 121.03	3.67 ± 7.94%	1373.95 ± 124.98	6.43 ± 2.89%
Double Deviant	85.99 ± 7.85	4.45 ± 8.72%	79.62 ± 8.98	7.49 ± 4.90%
Frequency Deviant	87.00 ± 7.29	3.33 ± 8.10%	81.38 ± 7.93	6.45 ± 4.56%
Duration Deviant	86.69 ± 7.59	3.68 ± 8.43%	80.73 ± 8.64	6.97 ± 4.35%

Mean ± Standard Deviation

Table 3.

Fronto-central 6 electrode mean proportion of variance for each Variance Component

Component	Peak		Mean		Window		Latency	
	Sorted	Traditional	Sorted	Traditional	Sorted	Traditional	Sorted	Traditional
Person (P)	35.24%	36.62%	36.05%	37.35%	31.19%	31.04%	15.48%	16.63%
Site (S)	0.03%	0.02%	0.05%	0.02%	0.38%	0.15%	2.42%	2.30%
Occasion (O)	0.09%	0.30%	0.10%	0.36%	0.36%	0.95%	0.15%	0.14%
P x S	21.77%	22.53%	21.16%	22.27%	15.99%	18.43%	26.75%	27.50%
P x O	1.35%	1.47%	1.24%	1.40%	1.98%	1.84%	1.39%	2.12%
S x O	0.92%	0.85%	0.90%	0.91%	0.79%	0.44%	1.27%	0.78%
Error + P x S x O	40.59%	38.21%	40.50%	37.70%	49.30%	47.15%	52.53%	50.54%
Relative Variance	98.96%	98.83%	98.96%	98.71%	98.47%	98.46%	96.15%	96.79%
G-coefficient	0.3562	0.3705	0.3644	0.3783	0.3168	0.3152	0.1610	0.1718

Table 4.

Fronto-central 6 electrode mean G-coefficeints by Measure, Deviant, and Averaging approach

Deviant	Peak		Mean		Window		Latency	
	Sorted	Traditional	Sorted	Traditional	Sorted	Traditional	Sorted	Traditional
DBL	0.34	0.31	0.36	0.32	0.38	0.34	0.16	0.15
DUR	0.48	0.4	0.49	0.51	0.41	0.43	0.16	0.22
FRQ	0.25	0.3	0.25	0.3	0.18	0.19	16	0.14

Table 5.

Fronto-central 6 electrode mean G-coefficients for Window measure by Site, Deviant, and Averaging approach

Sorted Averaging					Traditional Averaging				
Site	**DBL**	**DUR**	**FRQ**	**Grand Total**	**Site**	**DBL**	**DUR**	**FRQ**	**Grand Total**
Calgary	0.62	0.51	0.31	0.48	Calgary	0.47	0.29	0.47	0.41
Emory	0.30	0.61	0.42	0.45	Emory	0.66	0.77	0.39	0.61
Harvard	0.72	0.57	0.10	0.46	Harvard	0.48	0.58	0.38	0.48
Hillside	0.18	0.58	0.21	0.33	Hillside	0.39	0.38	0.29	0.35
Home	0.64	0.86	0.49	0.66	Home	0.42	0.82	0.51	0.58
UCLA	0.38	0.54	0.13	0.35	UCLA	0.47	0.52	0.07	0.35
UCSD	0.63	0.83	0.37	0.61	UCSD	0.48	0.83	0.35	0.56
UNC	0.28	0.41	0.51	0.40	UNC	0.22	0.39	0.34	0.32
Yale	0.62	0.35	0.45	0.47	Yale	0.65	0.37	0.47	0.50
Grand Total	**0.49**	**0.59**	**0.33**	**0.47**	**Grand Total**	**0.47**	**0.55**	**0.36**	**0.46**