UCSF

Title
Stability of mismatch negativity event-related potentials in a multisite study.

Permalink
https://escholarship.org/uc/item/4pj7w0sp

Journal
International journal of methods in psychiatric research, 29(2)

ISSN
1049-8931

Authors
Roach, Brian J
Hamilton, Holly K
Bachman, Peter
et al.

Publication Date
2020-06-01

DOI
10.1002/mpr.1819

Peer reviewed

ORIGINAL ARTICLE

Stability of mismatch negativity event-related potentials in a multisite study

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

[1]Department of Psychiatry, San Francisco Veterans Affairs Healthcare System, San Francisco, California

[2]Department of Psychiatry, University of California, San Francisco, California

[3]Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania

[4]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

[5]Division of Psychiatry Research, The Zucker Hillside Hospital, North Shore-Long Island Jewish Health System, Glen Oaks, New York

[6]Center for Psychiatric Neuroscience, Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York

[7]Department of Psychiatry, Hofstra North Shore-LIJ School of Medicine, Hempstead, New York

[8]Department of Psychiatry, Atlanta Veterans Affairs Medical Center, Decatur, Georgia

[9]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, Georgia

[10]Department of Psychiatry, Yale University, School of Medicine, New Haven, Connecticut

[11]Department of Psychiatry, University of California, San Diego, California

[12]Department of Psychiatry, Veterans Affairs San Diego Healthcare System, San Diego, California

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

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

[15]Semel Institute for Neuroscience and Human Behavior and Department of Psychology, University of California, Los Angeles, California

[16]Department of Psychology, Yale University, School of Medicine, New Haven, Connecticut

[17]Department of Molecular Medicine, Hofstra North Shore-LIJ School of Medicine, Hempstead, New York

[18]Department of Psychology, Emory University, Atlanta, Georgia

Correspondence
Daniel H. Mathalon, San Francisco VA Healthcare System/ Psychiatry Service (116D) 4150 Clement Street San Francisco, CA 94121, USA. Email: daniel.mathalon@ucsf.edu

Abstract

Objectives: Mismatch negativity (MMN), an auditory event-related potential sensitive to deviance detection, is smaller in schizophrenia and psychosis risk. In a multisite study, a regression approach to account for effects of site and age (12–35 years) was evaluated alongside the one-year stability of MMN.

Methods: Stability of frequency, duration, and frequency + duration (double) deviant MMN was assessed in 167 healthy subjects, tested on two occasions, separated by

52 weeks, at one of eight sites. Linear regression models predicting MMN with age and site were validated and used to derive standardized MMN z-scores. Variance components estimated for MMN amplitude and latency measures were used to calculate Generalizability (G) coefficients within each site to assess MMN stability. Trait-like aspects of MMN were captured by averaging across occasions and correlated with subject traits.

Results: Age and site accounted for less than 7% of MMN variance. G-coefficients calculated at electrode Fz were stable (G = 0.63) across deviants and sites for amplitude measured in a fixed window, but not for latency (G = 0.37). Frequency deviant MMN z-scores averaged across tests negatively correlated with averaged global assessment of functioning.

Conclusion: MMN amplitude is stable and can be standardized to facilitate longitudinal multisite studies of patients and clinical features.

KEYWORDS

event-related potential (ERP), generalizability, mismatch negativity (MMN), stability, standardization

1 | INTRODUCTION

Mismatch negativity (MMN) is an auditory event-related potential (ERP) component that is automatically evoked by an infrequently occurring "deviant" auditory stimulus that differs in duration, pitch, or another physical feature from a series of repeated preceding "standard" stimuli. MMN can be measured using either electroencephalography (EEG) or magnetoencephalography. It is believed to reflect sensory echoic memory, as detecting auditory deviance requires online formation and maintenance of a memory trace of immediately preceding standard stimuli. Because of its robust sensitivity to the pathophysiology of schizophrenia (Avissar et al., 2018; Erickson, Ruffle, & Gold, 2016), and its ability to predict conversion to a psychotic disorder in clinical high-risk (CHR) individuals (Bodatsch et al., 2011; Perez et al., 2014; Shaikh et al., 2012), MMN has great potential as an ERP biomarker in schizophrenia research, leading to its inclusion in the multisite North American Prodrome Longitudinal Study [NAPLS Addington et al. (2012)]. Accordingly, the test–retest reliability and stability of the MMN response over repeat test occasions must be evaluated to better understand the generalizability of this ERP component in multisite, clinical trials or longitudinal studies of psychosis.

The test–retest reliability of the MMN response has been the focus of several studies because of its potential clinical utility [see Naatanen (2003)] for a review), but reliability was assessed with Pearson (Kathmann, Frodl-Bauch, & Hegerl, 1999; Kujala, Kallio, Tervaniemi, & Naatanen, 2001; Pekkonen, Rinne, & Naatanen, 1995; Schroger, Giard, & Wolff, 2000; Tervaniemi et al., 1999; Uwer & von Suchodoletz, 2000) or Spearman (Deouell & Bentin, 1998; Schall, Catts, Karayanidis, & Ward, 1999) correlation coefficients in many of these studies. Alternatively, a measure that better captures agreement in responses from one test occasion to the next is the intraclass

correlation (ICC) coefficient (Shrout & Fleiss, 1979). ICCs have been calculated in some MMN studies (Biagianti et al., 2017; Chen, Chan, & Cheng, 2018; Hall et al., 2006; Lew, Gray, & Poole, 2007; Light et al., 2012; Light & Braff, 2005; McCleery et al., 2019; Recasens & Uhlhaas, 2017). Regardless of what coefficient type was reported, only three studies (Biagianti et al., 2017; Light et al., 2012; Light & Braff, 2005) had sufficiently long time-intervals between tests (at least 6 months) to be considered relevant to MMN stability.

Given the broad age range (12–35 years) in NAPLS (Addington et al., 2012), and potential age differences between CHR individuals who later transition to psychosis and those who do not, such studies must control for potential confounding effects of normal aging on MMN responses. One approach to adjust for any normal aging effects on MMN is to apply a simple linear regression model to the healthy control (HC) data. The resulting regression equation is used to calculate age-corrected MMN z-scores for all subjects. This is done by subtracting the predicted MMN based on a subject's age from his/her observed MMN score, and then dividing the result by the standard error of regression obtained from the HC age-regression model. Such age-corrected MMN z-scores derived from a HC model have no relationship with age in the HC sample, but any pathological age effects in patient or hold-out samples are preserved. The z-scores are readily interpretable as linear transformations of MMN raw scores, reflecting the degree of deviation or abnormality, in standard units, from the MMN expected for a person of a given age in the HC sample. This approach is not unprecedented, having been implemented previously in MMN (Biagianti et al., 2017; Perez et al., 2014), other ERP (Hamilton, Roach, et al., 2019; Hamilton, Woods, et al., 2019; Mathalon et al., 2018; Mathews et al., 2016; Perez, Woods, et al., 2012; Perez, Ford, et al., 2012), functional (Fryer et al., 2013; Fryer et al., 2016; Fryer et al., 2018) and structural

(Heyes et al., 2001; Jernigan, Press, & Hesselink, 1990; Pfefferbaum et al., 1992) magnetic resonance imaging studies.

In any multisite study design comparing patients and controls, the effect of laboratory testing site should be carefully considered (Glover et al., 2012). Despite applying all possible best practices to minimize site-specific influences on experiments, differences in mean responses between sites (i.e., fixed effects of site) remain possible due to real differences in the random samples of participants studied at each site. Therefore, the site effect should be modeled and, even when site is not statistically significant, should not be disregarded based on this criteria alone. Furthermore, it may be of interest to test secondary/exploratory hypotheses that do not involve the entire study sample. In such cases, simply including site in the model may not be sufficient if the subset of participants is too small to accurately estimate site effects.

In NAPLS, the effect of site can be added as a categorical covariate in the age model. In this new model, a common age effect is estimated across all sites such that site effects on either MMN or age cannot create a spurious relationship between MMN and age. Moreover, the error from this new model can be used to calculate site- and age-adjusted MMN z-scores, reflecting the difference, in standard units, from the MMN expected for a person of a given age, *from a particular site* in the HC sample. Such z-scoring is particularly useful for planned comparisons between CHR individuals who convert (CHR-C) to a psychotic disorder and those who do not (CHR-NC) within 24 months of initial NAPLS baseline assessments because there is no feasible a priori method to match these subjects on age and/or sample an equal number from each site.

Accordingly, the goals of this study were to (a) describe a set of site and age regression models in HCs that will be used to create standardized, site- and age-adjusted MMN z-scores for all NAPLS subjects, (b) estimate variance components and associated G-coefficients representing the single site stability of MMN responses measured with different scoring methods separately at each of the eight NAPLS sites, and (c) compare such G-coefficients calculated using raw and z-scored MMN responses. Additional exploratory analyses are presented to demonstrate the Spearman-Brown prophecy (Brown, 1910; Spearman, 1910) in practice by averaging across MMN measured on separate occasions to capture more trait-like aspects of the MMN and relate it to subject traits.

2 | METHODS

2.1 | Participants

Participants were recruited at each of the eight NAPLS2 sites and all provided written, informed consent to participate in this IRB-approved study. EEG data were collected at baseline assessment from 241 HCs, and 167 (~70%) of these HCs completed at least one follow up EEG assessment. Additional demographic characteristics of these 167 subjects are presented in Table 1.

All HCs had at least one global assessment of functioning (GAF) as a part of study procedures (Endicott, Spitzer, Fleiss, & Cohen, 1976; Jones, Thornicroft, Coffey, & Dunn, 1995). The current GAF score nearest to the baseline EEG date (median time between GAF and EEG was 22 days, IQR: 7–34.61 days) was saved for correlation analyses. As an additional GAF metric, the mean current GAF score across all assessments (max = 5, one every 6 months) during the 24 month study period was saved. There were 18 HCs (~11%) who only had one GAF assessment, making their baseline GAF and mean GAF scores equivalent.

2.2 | MMN paradigm

All sites used similar hardware and presentation software (www.neurobs.com) to run the EEG experiment. Auditory stimuli were delivered via ER1-A Etymotic insert earphones and subjects responded with a Cedrus RB-830 button box. 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

TABLE 1 Demographic Information

Site	Subjects Number	Gender (M, F)	Test age (mean ± SD)	Re-test age (mean ± SD)	Education (mean ± SD)	Days between tests		
						Min	Median	Max
UCLA	21	10, 11	18.15 ± 3.05	19.22 ± 2.98	11.38 ± 3.01	217	380	602
Emory	20	13, 7	21.82 ± 5.03	22.88 ± 5.03	13.6 ± 3.69	145	354	795
Harvard	23	11, 12	18.9 ± 4.56	19.94 ± 4.45	11.26 ± 3.15	77	353	686
Hillside	23	15, 8	17.1 ± 2.73	18.38 ± 2.72	10.78 ± 2.61	270	381	812
UNC	23	14, 9	20.3 ± 2.52	21.3 ± 2.51	13.78 ± 2.28	272	364	462
UCSD	17	13, 4	20.38 ± 6.64	21.24 ± 6.62	12.59 ± 4.05	147	350	400
Calgary	25	9, 16	21.76 ± 5.89	22.79 ± 5.81	13.6 ± 4.53	255	363	729
Yale	15	6, 9	21.51 ± 6.13	22.56 ± 6.08	12.8 ± 4.31	254	370	762
TOTAL	**167**	**91, 76**	**19.91 ± 4.89**	**20.97 ± 4.83**	**12.46 ± 3.61**	**77**	**365**	**812**

(DBL) presented for 100 ms at 1000 Hz. A total of 1,794 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 MMN, participants were instructed to ignore auditory stimuli and focus on a separate distractor task. A visual oddball paradigm was run simultaneously with MMN, where image presentation was jittered to avoid cooccurring visual oddball and MMN ERPs.

2.3 | EEG data acquisition

EEG was digitized at 1024 Hz using 32- or 64-channel electrode caps (Biosemi, Amsterdam, The Netherlands), and the common 32 channels were used in subsequent steps. Additional electrodes were placed on the above and below the right eye, on the outer canthus of each eye, and on the mastoids. An offline average mastoid reference was initially applied to continuous EEG data prior to all preprocessing.

2.4 | Preprocessing

Mastoid-referenced, continuous EEG recordings were high-pass filtered at 1 Hz prior to segmentation into 1,000 ms epochs (−500 to 500 ms). Blinks and eye movement artifacts were recorded by electrodes placed around the eyes and were subtracted from single trials using regression (Gratton, Coles, & Donchin, 1983). Following baseline correction (−50 to 0 ms), outlier electrodes were interpolated within single trial epochs based on previously established criteria (Nolan, Whelan, & Reilly, 2010). A spherical spline interpolation (Delorme & 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 were rejected, if they contained amplitudes greater than ±100 μV in any of these electrodes: AF3, AF4, F3, Fz, F4, FC1, FC2, FC5, FC6, C3, Cz, C4.

2.5 | ERP averaging and MMN measurement

ERP averages for all stimulus types were determined using a sorted averaging method, which 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 for each subject (Rahne, von Specht, & Muhler, 2008). In this study, single-epoch root mean squared (RMS) amplitude values averaged across the 12 electrodes used for artifact rejection for each trial are calculated and sorted in ascending order for each stimulus type. Following averaging, ERPs for all stimulus types were low-pass filtered at 30 Hz, and then standard tone ERPs were subtracted from deviant ERPs to obtain difference waves. MMN peak amplitude was classified as the most negative peak

between 90 and 290 ms in the difference wave. MMN mean amplitude ±10 ms around the peak was also quantified as an alternative measurement to peak amplitude. Average amplitude in a fixed window defined based on grand average waveforms (90–170 ms for FRQ and DBL, 150–230 for DUR) was quantified as a third approach. Peak latencies were saved for a fourth set of analyses.

2.6 | Common slope linear regression models

When experimental data come from one laboratory site, an ordinary least squares (OLS) regression model can be applied to MMN (or other response variable) data using age as a predictor to obtain a simple linear equation that can be used to predict a subject's MMN response at a given age. Such an OLS model predicting MMN scores by age may have the form:

$$y_i = \beta_0 + a_i * \beta_0 + e_i \tag{1}$$

In Equation (1), y is the MMN score, a is the age, and e is the residual (i.e., difference between age-predicted and actual MMN score) for the ith subject. The model error, or specifically, root-mean-square error (RMSE), is calculated as:
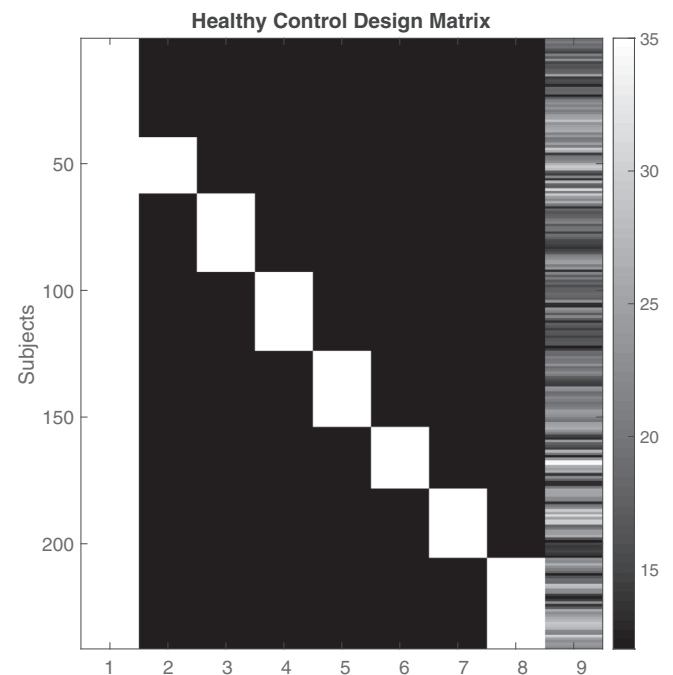

FIGURE 1 A graphical representation of the design matrix used to predict mismatch negativity (MMN) responses is plotted. The first eight columns include an intercept term and seven site indicator variables, where white represents 1 and black 0. These columns capture fixed effects of site, while the ninth column is the age covariate, with grayscale age value representing each of the 241 healthy control participants' ages at baseline MMN assessment. This model is the common slope model where each site may have a different y-intercept, but a common age relationship estimated using data from all subjects and sites

$$RMSE = \sqrt{\frac{\sum_{i}^{n}(e_i{}^2)}{n}} \qquad (2)$$

The RMSE summarizes the model's error across all subjects, and it can be used to calculate an age-corrected MMN z-score:

$$z_i = \frac{e_i}{RMSE} \qquad (3)$$

In the multisite setting, one must consider laboratory site as between subjects, categorical variable. This increases the OLS model design matrix from two (intercept + age) to nine columns in NAPLS. The additional seven columns are indicator variables that capture site
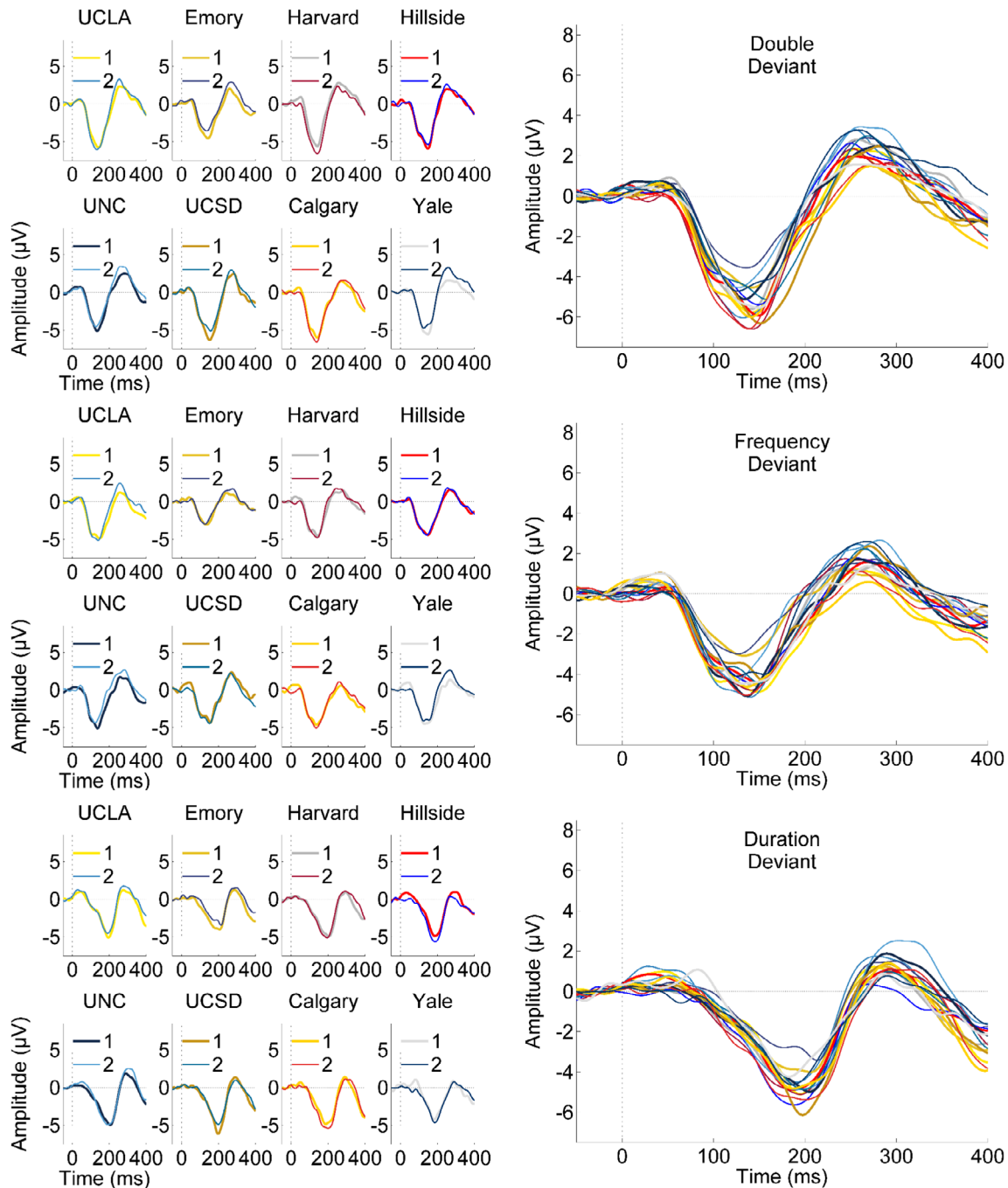
FIGURE 2 Site- and session-specific grand average mismatch negativity (MMN) deviant minus standard tone difference waveforms are plotted for the Double (Frequency plus Duration) Deviant (Top), Frequency Deviant (Middle), and Duration Deviant (Bottom) from electrode Fz. Grand Average MMN waveforms for each NAPLS laboratory site are plotted separately on the left-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 right-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

membership (i.e., 1 if a subject is from that site, 0 otherwise), and only seven indicator variables are needed to encode eight sites. The corresponding design matrix for all 241 HCs included in the baseline analysis is plotted in Figure 1.

This same design matrix is applied to all response variables (MMN amplitude, latency, etc.), and the resulting parameter estimates are used to obtain expected responses for a particular subject given subject age and site. The difference between a given subject's actual value and that predicted value, divided by the RMSE [Equation (2)] of the model yields an age- and site-corrected z-score, which represents that subject's deviation, in standardized units, from the expected value for a subject who is the same age, measured at the same site.

In addition to the standard assumptions of a regression model, this design assumes (a) the age relationship does not differ between sites, and (b) there is not a higher order polynomial (e.g., quadratic) age relationship with the response variable. Both of these assumptions can be formally tested by either (a) adding site*age interaction effects to the model or (b) adding a mean-centered, age-squared term to the model and checking for a statistically significant improvement in model fit with the r^2 change F-test. In the age-squared case, this is equivalent to the test of the relationship between the response variable and the age-squared term. In the more complicated site*age model, heterogeneous age relationships at the sites would lead to an improved fit. Such F-tests were conducted for all variables (384 total), and both Akaike and Bayseian-Swartz Information Criteria (AIC and BIC) were calculated as additional descriptive measures of model fitness (Sakamoto & Kitagawa, 1987). False discovery rate (FDR) correction was applied separately to the two sets of F-tests, and Bonferroni correction was separately applied to families of tests limited to the electrodes (n = 32) for each measure and deviant type (p = 0.05/32 = 0.0015625). Finally, the number of uncorrected (p < .05) significant tests was listed for descriptive purposes.

Follow-up longitudinal MMN data were also z-scored using the baseline HC model and t- or F-tests were conducted to assess age and site effects, respectively, as additional measures of model fitness. This subset of 167 follow up data points could be considered a "hold-out" data set, and any site or age effects indicate that the z-scoring procedure suboptimally accounted for linear effects of age and fixed effects of site.

2.7 | Variance components and G-coefficients

The longitudinal HC data were re-purposed as a single facet (test occasion) G-study design to estimate variance components. Such a design allows estimation of three variance components for any response using the data from the participants at a particular site. The variance components for Person (σ_p^2), Occasion (σ_o^2), and Person x Occasion plus Error (σ_{po+e}^2) are estimated separately for each NAPLS laboratory site, as Site may represent another source of variance [see Roach et al. (2019)]. Once variance components are estimated, the G-coefficient, which provides a measure of generalizability or stability of the measured score in this longitudinal setup, can be calculated as in Equation (4):

$$G = \frac{\sigma_p^2}{\left(\sigma_p^2 + \frac{\sigma_{po+e}^2}{n_o} \right)} \quad (4)$$

The NAPLS2 study design included EEG assessments at baseline, 12 month, and 24 month study time points. MMN scores from each session are treated separately, with particular emphasis on using baseline data to predict conversion to psychosis, meaning the best choice for n_o is 1. Therefore, the G-coefficient is equal to the intraclass correlation (ICC) defined by Shrout and Fleiss (e.g., ICC (3,1) in (Shrout & Fleiss, 1979)) when n_o = 1. Variance components were estimated using a restricted maximum likelihood approach in Matlab (Witkovský, 2012). Components were estimated separately and saved for the three deviant types (DBL, FRQ, DUR), 32 electrodes, and four MMN measurements (peak amplitude, mean around peak, mean in fixed window, and peak latency) for both MMN raw scores and z-scores.

The goal of a G-study is not to test a specific hypothesis. Thus, there are no p-values associated with estimated variance components or

TABLE 2 Trial Numbers in ERP averages

Trial type	Test	Re-test
Standard	1,356.04 ± 108.13	1,329.51 ± 157.72
Double deviant	80.37 ± 7.42	78.43 ± 10.64
Frequency deviant	79.99 ± 7.06	79.28 ± 9.88
Duration deviant	80.26 ± 6.98	78.93 ± 9.93

Note: Mean ± Standard Deviation.
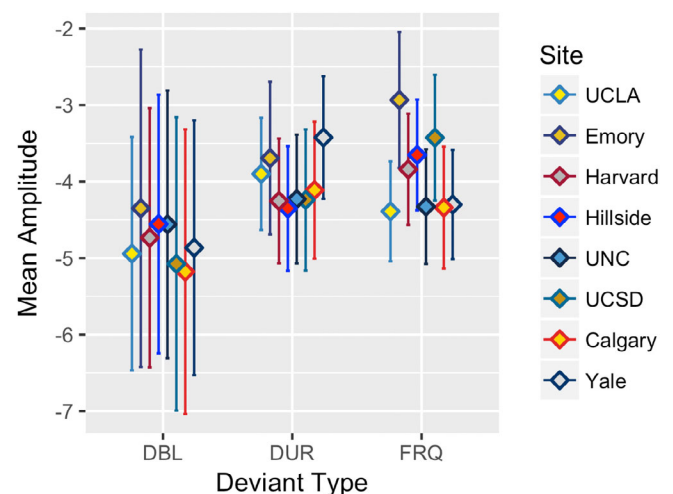
FIGURE 3 Estimated 95% confidence intervals are plotted for the mean mismatch negativity (MMN) amplitude at 18 years of age. The double (frequency plus duration; DBL, left), frequency (FRQ, middle), and duration (DUR, right) deviants are plotted separately along the x-axis from electrode Fz, and are separated and color-coded by NAPLS laboratory site. Estimates were derived from a site and age regression model of MMN amplitude averaged across a fixed window of either 90–170 ms (for DBL and FRQ) or 150–230 ms (DUR)
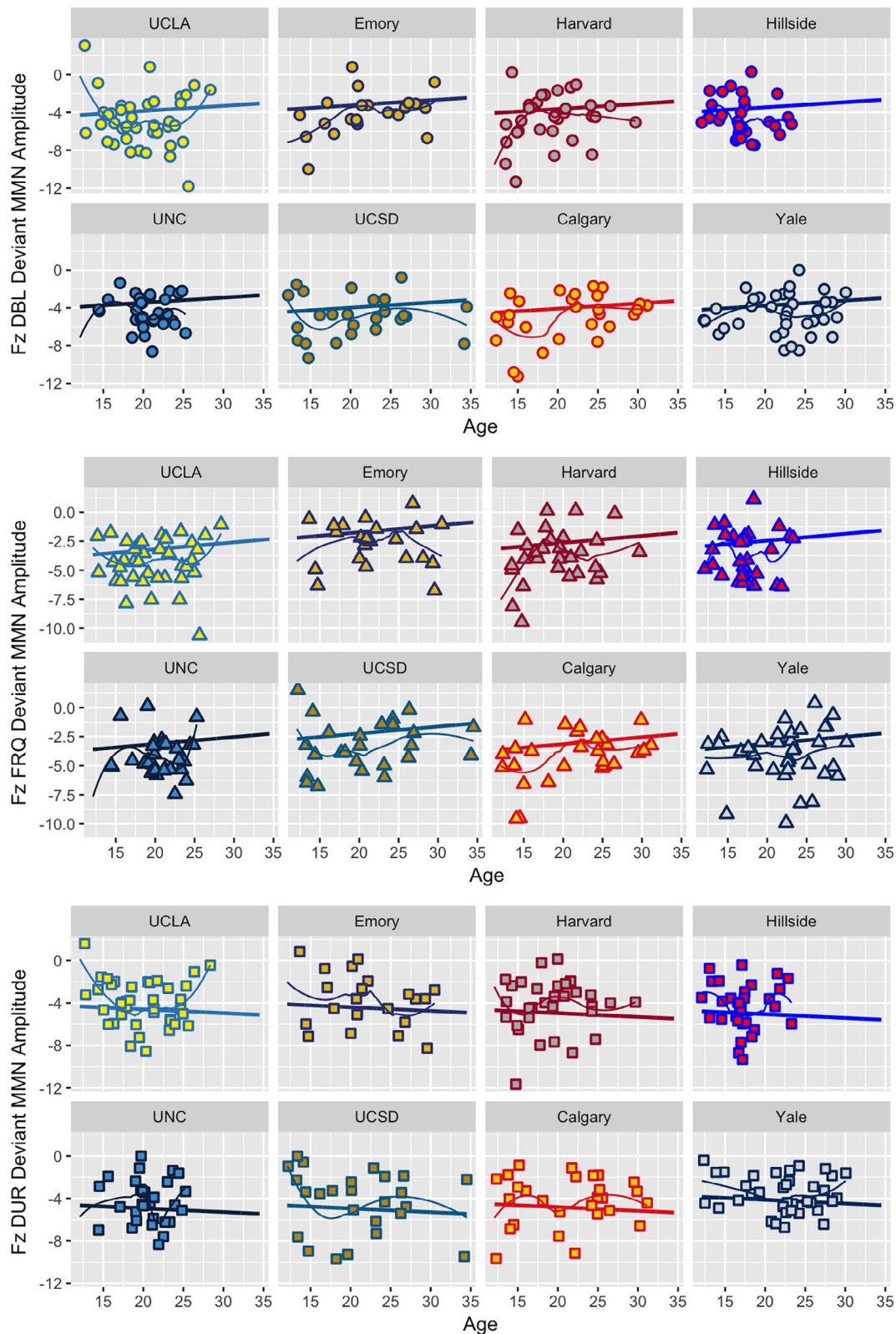
FIGURE 4 Scatterplots depict the relationships between mismatch negativity (MMN) amplitude averaged across a fixed window at electrode Fz and participant years of age at testing for double (frequency plus duration; DBL, circles), frequency (FRQ, triangles), and duration (DUR, squares) deviants. Data are plotted separately for each site, and thick lines depict the common age relationship across sites based on regression models. Thin lines depict site-specific, nonlinear locally weighted predictions of MMN given age (Cleveland, Grosse, & Shyu, 1992), and there is no higher-order polynomial or other nonlinear pattern of fit that is consistent across sites

G-coefficients. However, existing guidelines for determining clinical significance of ICCs suggest that the reliability coefficient can be qualitatively categorized as follows: ICC < 0.4 is poor, 0.4 ≤ ICC < 0.6 is fair, 0.6 ≤ ICC < 0.75 is good, and 0.75 ≤ ICC < 1 is excellent (Cicchetti & Sparrow, 1981). Therefore, G-coefficients were categorized using these 4 labels for descriptive purposes, as done previously (Roach et al., 2019).

2.8 | Exploratory correlations between MMN variables and trait variables

To capture more trait-like aspects of MMN, z-scores from 6 fronto-central electrodes (F3, Fz, F4, C3, Cz, C4) were averaged across the two test occasions separately for each deviant type and correlated with mean GAF or used to explore gender differences in MMN. As one method to demonstrate the enhanced reliability of averaged MMN z-scores, baseline MMN z-scores, and baseline GAF scores were also correlated. Similar to age regression models, site was a categorical covariate and heterogeneity of MMN-GAF relationships between sites were ruled out by first including a site*GAF interaction term, and the r^2 change F-test was used to determine improvement in model fit.

Given the exploratory nature of these correlations, parameter estimates, uncorrected p-values, as well as FDR-corrected p-values within this trait family of tests are reported.

3 | RESULTS

MMN ERP waveforms from electrode Fz are plotted in Figure 2. There is consistency between waveforms at each site and on each test occasion despite the long interval between tests and differences in site demographics. Descriptive statistics for trial numbers contributing to individual subject and test occasion ERPs are included in Table 2.

3.1 | Common slope linear regression models

There were two sets of F-tests to assess the appropriateness of common slope, site, and age regression models. In one set, a mean-centered age-squared term was added to the model to test for quadratic age relationships with MMN scores. Only 7.8% (30/384) of these tests showed statistically significant quadratic age relationships at an uncorrected level (ps < 0.05), none survived electrode-family Bonferroni-correction (all ps > 0.0015625), and none were significant after FDR-correction. Comparisons of AIC and BIC between age and age-squared models indicated that the age-squared model was better (i.e., smaller AIC or BIC values) for ~25% (95/384) of the models based on AIC but only 4% (16/384) based on BIC. This indicates that a quadratic age effect does not systematically improve MMN modeling and should be omitted.

In the second set of F-tests, age*site interaction effects were added to the model to determine if there were site-specific differences in MMN-age relationships. Only 4% (16/384) of age*site F-tests

showed evidence of uncorrected effects (ps < 0.05), one survived Bonferroni-correction (p < 0.0015625) and none were significant after FDR correction. This more complicated model was better than the simplified model based on AIC in 5.5% (21/384) of the models and none of the models for BIC. This indicates that the age relationship did not systematically differ between the sites.

In the common slope models, 33.3% (128/384) of the tests of age relationships were statistically significant at an uncorrected level (ps < 0.05), 13% (50/384) survived Bonferroni-correction (ps < 0.0015625), and 18.5% (71/384) survived FDR correction.
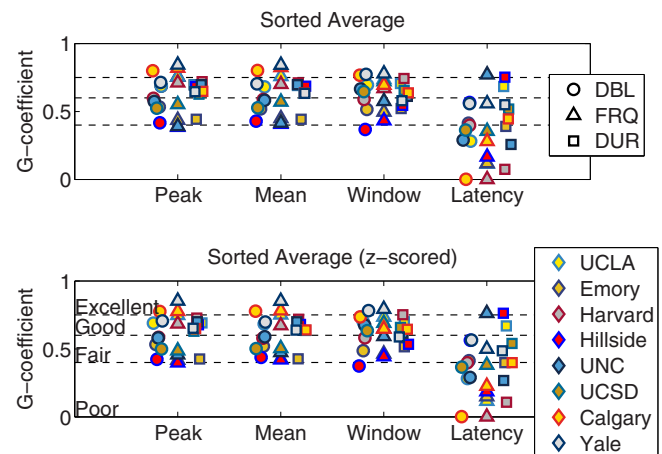
FIGURE 5 G-coefficients for the single-facet (test occasion) generalizability substudies calculated separately for each NAPLS geographic site for electrode Fz based on either raw (top) or standardized mismatch negativity (MMN) z-scores (bottom). Measurement approaches are plotted along the x-axis separately for double-deviant (DBL, circles), frequency-deviant (FRQ, triangles), and duration-deviant (DUR, squares) mismatch negativity. These include peak amplitude ("Peak": most negative peak between 90 and 290 ms in the MMN difference wave), mean amplitude ("Mean": ±10 ms around the peak), average amplitude in a fixed window ("Window": 90–170 ms for FRQ and DBL, 150–230 for DUR), and peak latency ("Latency"). Dashed lines indicate qualitative categorization of G-coefficients based on preexisting standards (Cicchetti & Sparrow, 1981)

TABLE 3 Frequency of G-coefficients by NAPLS Site

Site	Poor	Fair	Good	Excellent	Total
UCLA	136	99	117	32	384
Emory	168	192	23	1	384
Harvard	182	119	72	11	384
Hillside	197	135	46	6	384
UNC	169	114	61	40	384
UCSD	141	106	116	21	384
Calgary	130	103	104	47	384
Yale	108	146	77	53	384
TOTAL	1,231	1,014	616	211	3,072
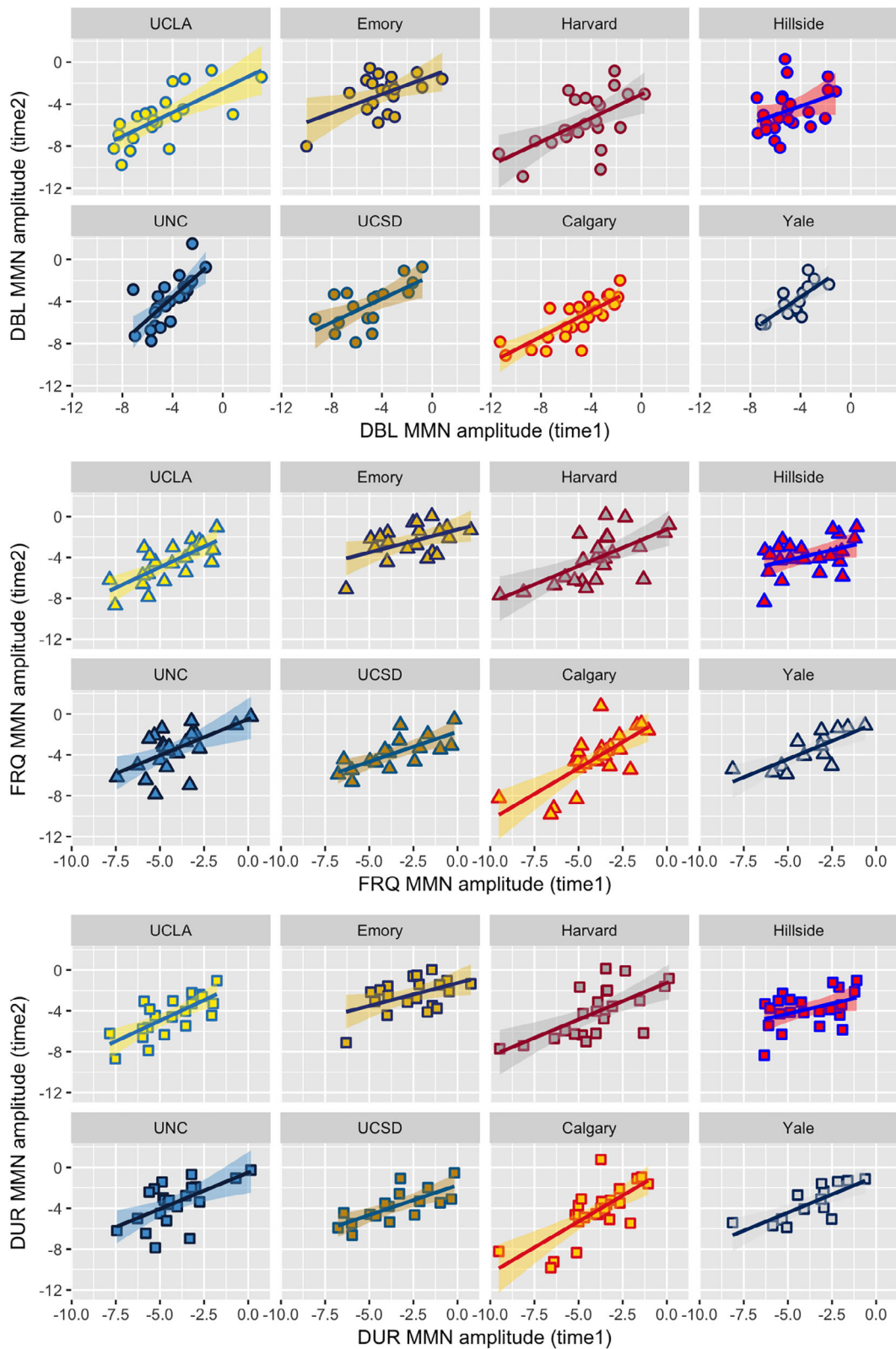
FIGURE 6   Scatterplots depict the relationships between mismatch negativity (MMN) amplitude averaged across a fixed window at electrode Fz at first (Time 1, x-axis) and second (Time 2, y-axis) test occasions for double (frequency plus duration; DBL, circles), frequency (FRQ, triangles), and duration (DUR, squares) deviants. Data are plotted separately for each site, and thick lines depict the site-specific linear relationship between occasions along with shading to show 95% confidence intervals

Deviant	Term	Estimate	S.E.	t-statistic	p-value	FDR p-value
DUR	Male vs female	0.032	0.146	0.218	0.82745	1.00000
FRQ	Male vs female	−0.049	0.130	−0.379	0.70534	1.00000
DBL	Male vs female	0.097	0.132	0.730	0.46647	1.00000
DUR	Mean GAF score	−0.011	0.009	−1.196	0.23350	
FRQ	Mean GAF score	−0.021	0.008	−2.679	0.00816	0.04078
DBL	Mean GAF score	−0.017	0.008	−2.140	0.03386	0.13542

TABLE 4 Parameter estimates for trait-like average MMN exploratory models

Across all models, site and age accounted for 6.32% of MMN variance (range: 0.2163–14.2434%), indicating that even for the strongest site and age effects, at least 85% of the variance in the MMN raw scores remained in the site- and age-corrected MMN z-scores. Using these regression models, site-specific mean MMN amplitude for the window measure from electrode Fz and 95% confidence intervals (CIs) for an 18 year-old subject were estimated and plotted in Figure 3.

The estimated means and CIs demonstrate that there is overlap between the FRQ and DUR MMN amplitude across sites, with the FRQ MMN being smallest at Emory. The DBL MMN estimates are slightly larger (i.e., more negative MMN amplitudes) than the other deviants with CIs that are approximately twice as wide as those for the other deviants. The common age effect is plotted on top of site- and deviant-specific scatter plots in Figure 4 for electrode Fz.

Tests of age relationships in the z-scored longitudinal follow-up MMN data indicated that the age effect was removed in this hold-out subsample, with only ~5% (20/384) statistically significant effects at an uncorrected level (p < 0.05), consistent with what is expected by chance. None survived FDR correction. Tests of site effects in these data indicated ~15% (58/384) were significant at an uncorrected level. Only 3 site tests of MMN latency measures survived FDR correction.

3.2 | Variance components and G-coefficients

G-coefficients for each electrode, deviant type, measure, and NAPLS site are included in Table S1 for both raw MMN and z-scores. As can be seen in Figure 5, the G-coefficients for Fz are fair or better (G ≥ 0.4) in almost every MMN amplitude measure across NAPLS sites, but the latency G-coefficients are highly variable and poor in many cases.

Frequencies of poor, fair, good, and excellent reliability categorization of all G-coefficients are presented in Table 3 separated by NAPLS site. The table demonstrates that the majority (~60%) of G-coefficients were fair or better, including many (~27%) scores with excellent generalizability. G-coefficients based on z-scores were nearly equivalent to those based on raw scores (average difference in G-coefficients = 0.0044), consistent with relatively small proportions of MMN raw score variance being accounted for by age and site. Site-specific relationships between mean MMN amplitude in a fixed window on the first and second test occasions are plotted along with corresponding deviant-specific scatter plots in Figure 6 for electrode Fz.

3.3 | Exploratory correlations between MMN variables and trait variables

Parameter estimates along with test statistics for all trait-like MMN models are presented in Table 4. There were no significant site*GAF interaction effects for either DBL (F[7,151] = 0.974, p = 0.45) or FRQ (F[7,151] = 0.867, p = 0.5341) MMN, but there was evidence of heterogeneous DUR MMN-GAF relationships between sites (F[7,151] = 2.1644, p = 0.0401, r^2 = 0.1185). Scatter plots of the relationships between each deviant type and mean GAF scores are plotted separately for each site in Figure 7. The plots show mostly negative relationships (i.e., greater GAF is associated with more negative MMN) for FRQ and DBL MMN, but only negative relationships between DUR MMN and GAF at Emory (t[18] = −2.545, p = 0.0203, r^2 = 0.2647) and UCSD (t[15] = −2.431, p = 0.0281, r^2 = 0.2826). Reduced models for DBL and FRQ MMN revealed negative relationships with GAF (DBL: r^2 = 0.0564, FRQ: r^2 = 0.0598), controlling for site, but only the FRQ MMN effect survived FDR correction (Table 4). Had trait-like aspects of MMN and GAF not been emphasized through averaging across assessments, neither the time 1 FRQ MMN ($\hat{\beta}$ = −0.011, t(158) = −1.5, p = 0.135, r^2 = 0.02) nor the time 1 DBL MMN ($\hat{\beta}$ = −0.001, t(158) = −0.177, p = 0.86, r^2 = 0.0072) z-score relationships with nearest current GAF score would have reached statistical significance.

In the gender models, there was neither evidence of a site*gender interaction effect for any deviant type (all ps > 0.487), nor evidence of a gender difference between males and females in the reduced models.

4 | DISCUSSION

One goal of this study was to present a site and age modeling strategy to create regression models to produce standardized site- and age-adjusted MMN z-scores for all participants and all test occasions in NAPLS2, and in doing so, demonstrate the utility of such an approach for large, multisite studies. The main purpose of the generalizability analyses presented was to quantify variance components and associated G-coefficients representing the single site, single session stability of MMN responses measured about 1 year apart. G-coefficients indicated that for both raw MMN and age and site- adjusted z-scores, the stability of amplitude measures was fair or better and consistent across the 8 laboratory sites, while the stability of latency measures
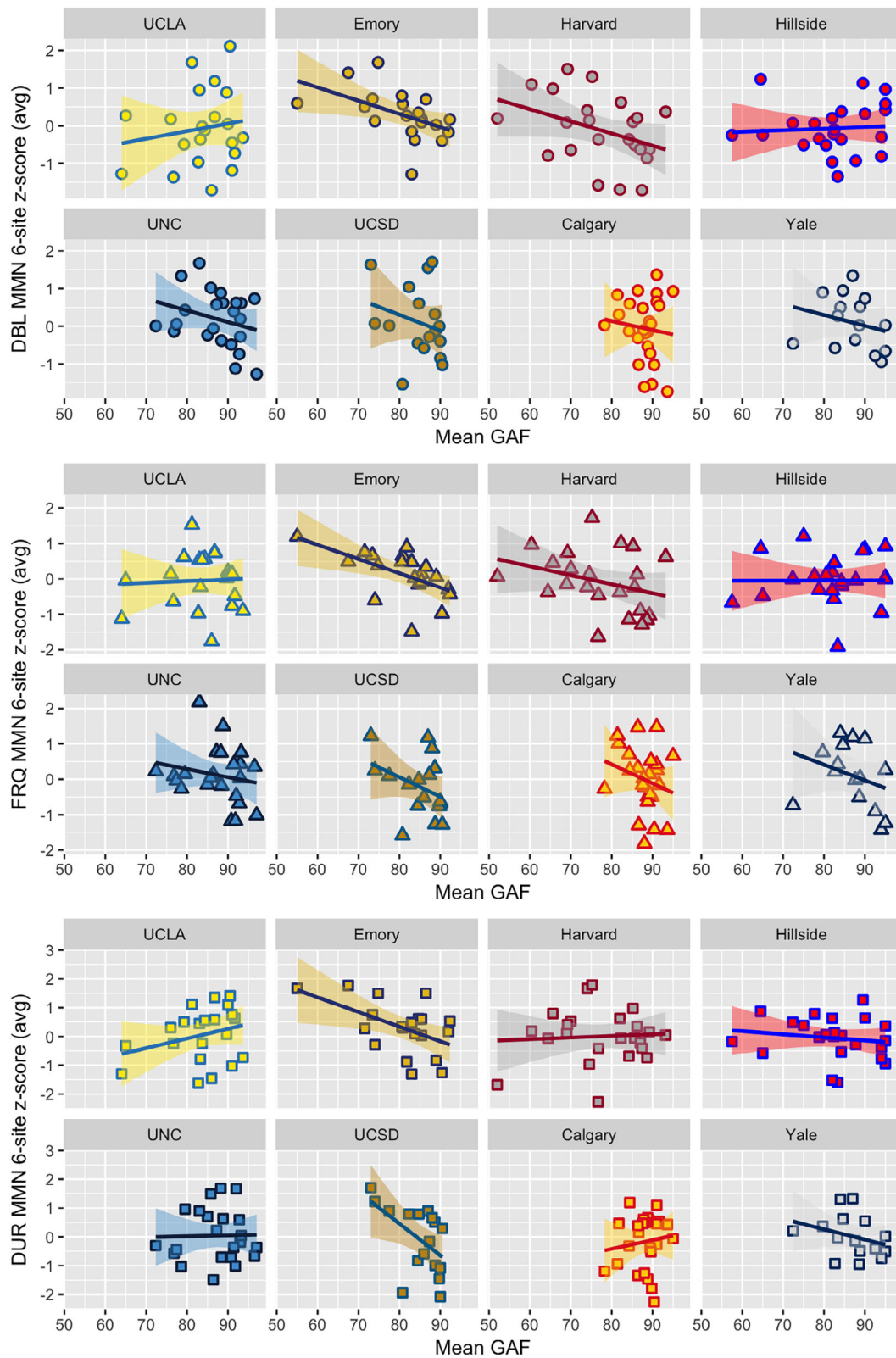
FIGURE 7 Scatterplots depict the relationships between mean scores for Global Assessment of Functioning (GAF) averaged across all study time points (x-axis) and standardized z-scores from mismatch negativity (MMN) amplitude averaged across a fixed window averaged across two test occasions and electrodes F3, Fz, F4, C3, Cz, C4 for double (frequency plus duration; DBL, circles), frequency (FRQ, triangles), and duration (DUR, squares) deviants (y-axis). Data are plotted separately for each site, and thick lines depict the site-specific linear relationship between mean MMN along with shading to show 95% confidence intervals. Most sites and deviants show negative relationships, indicating that better mean GAF scores are associated with larger (i.e., more negative) MMN z-scores

was inconsistent across sites and poor in many cases. This suggests that amplitude measures are optimal for longitudinal studies of MMN.

Several alternatives to the z-score approach for removing site- and age-related confounds in clinical studies are potentially problematic. One alternative is to ignore site- and age-related variation. A second approach would be to eliminate subjects from certain sites in order to match groups on age at each site. In studies of rare outcome events or patients, which is one of the motivations of a multisite study like NAPLS, eliminating subjects is disadvantageous. A third approach is to conduct ANCOVA with site and age as a covariates. The problem with an age factor in ANCOVA is that MMN-age relationships are derived from a pooled estimate of aging effects from all of the groups being compared, including the CHR. On theoretical grounds, it is reasonable to hypothesize that physiological measures like MMN in the psychosis prodrome may have abnormal age trajectories, reflecting abnormal brain maturation, and other pathogenic processes operating during the transition to psychosis or disease-related progressive brain changes occurring after illness onset (Kiang, Braff, Sprock, & Light, 2009; Light et al., 2015; Todd et al., 2008). Accordingly, we believe ANCOVA models are inappropriate because of their potential to remove disease-related aging effects along with normal aging effects within the study sample.

Two previous studies reported excellent MMN reliability (Fz ICC > 0.8) using a long duration deviant similar to this study and a window measurement (135–205 ms) from nose-referenced data (Light et al., 2012; Light & Braff, 2005). These reliability coefficients were based on either 10 patients with schizophrenia (Light & Braff, 2005), 168 patients with schizophrenia, or 58 healthy subjects (Light et al., 2012), tested twice, at least 1 year apart. While the corresponding window measure G-coefficients, averaged across all 8 NAPLS sites, was smaller in the present study (raw and z-score MMN at Fz G = 0.625), the subjects in this study were younger, healthy participants who may have experienced more true score change in a 1 year interval than the older schizophrenia patients and controls in other studies. A similar age group of 28 young, healthy participants (Biagianti et al., 2017) had good duration deviant peak amplitude reliability based on two MMN sessions, approximately 6 months apart (Fronto-central 6 electrode average ICC = 0.72), which is closer to reliability averaged across all NAPLS sites in this study (raw and z-score peak MMN at Fz G = 0.644) and consistent with the idea that more true score change occurs in younger subjects. It is also worth noting that when averaging all 8 NAPLS sites' separately calculated G-coefficients in our traveling subjects study, where subjects were tested on two consecutive days at each site, the duration deviant MMN based on the window measure similarly has ~60% of the relative variance attributed to persons, and 40% attributed to error, on average [raw MMN at Fz G = 0.6, Roach et al. (2019)]. These estimates are consistent with other MMN reliability studies of healthy subjects that also reported good reliability using long duration deviants (Fz ICC = 0.66 in Hall et al., (2006)] or frequency deviants [Cz ICC = 0.6 in Lew et al. (2007)].

The Spearman-Brown prophecy formula indicates that reliability of a score increases as test length or the number of items averaged to summarize a subject's score increases (Brown, 1910; Spearman, 1910). In the case of this study, averaging across the two EEG test occasions reduces the contribution of the error variance component to the calculation of the G-coefficient in Equation 4. This shifts the average G-coefficient at Fz for all deviant types from good (G >= .6) to excellent (G >= .75). In practice, this averaging emphasized the trait-like attributes of the MMN scores, allowing relationships between averaged GAF scores and FRQ MMN to emerge. This negative correlation between GAF and MMN has previously been observed in schizophrenia patients using DUR MMN (Fulham et al., 2014; Jahshan et al., 2012; Koshiyama et al., 2018; Light & Braff, 2005). Future studies exploring the relationship between MMN and functioning should consider averaging across multiple assessments to emphasize trait-like aspects of MMN and functioning measures while also reducing error variance. There were no gender differences in averaged MMN scores, consistent with some (Qiao et al., 2015; Yang et al., 2016) but not all (Light et al., 2015) prior reports.

There are several limitations in the present stability analyses that should be carefully considered. Because estimates of variance components can be fairly unstable when the number of measurements is small, having only a subset of all the HC subjects studied on only two test occasions at each site is not ideal. It is possible that HCs who returned for a second EEG assessment represent a biased subgroup of subjects who were above-average in compliance, leading to inflated G-coefficients. For example, the Yale site had the lowest number of subjects in their G-study, the lowest retention rate (41.67%), and the greatest number of excellent G-coefficients. However, the Calgary site had the most subjects, the best retention rate (92.6%), and the second greatest number of excellent G-coefficients.

Despite these limitations, MMN amplitude measures appear to have fair or better stability across all NAPLS sites, similar to the within-site test–retest reliability previously reported in a small (N = 8) sample traveling subjects study (Roach et al., 2019). Furthermore, site- and age-standardization of MMN measures via linear regression minimally changed the G-coefficients while removing fixed effects of site and age in the full (N = 241) NAPLS2 HC sample. These MMN z-scores can be used to test for pathological aging effects in the CHR sample and to test hypotheses in subsamples of subjects that may not be balanced in number and/or age across the 8 sites (e.g., comparing CHR-C to CHR-NC). This simple, linear transformation represents a useful approach to multisite EEG studies of rare patient populations or clinical trials. The consistency of MMN waveforms and G-coefficients across site between two test occasions indicates that MMN amplitude measures are generalizable, and like in other consortium studies (e.g., Light et al., 2015), it is feasible to combine data from multiple, appropriately controlled and calibrated, research laboratory sites to study MMN.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

ORCID

Brian J. Roach https://orcid.org/0000-0002-3264-1465

REFERENCES

Addington, J., Cadenhead, K. S., Cornblatt, B. A., Mathalon, D. H., McGlashan, T. H., Perkins, D. O., ... Cannon, T. D. (2012). North American Prodrome longitudinal study (NAPLS 2): Overview and recruitment. *Schizophrenia Research*, 142(1-3), 77–82. https://doi.org/10.1016/j.schres.2012.09.012

Avissar, M., Xie, S., Vail, B., Lopez-Calderon, J., Wang, Y., & Javitt, D. C. (2018). Meta-analysis of mismatch negativity to simple versus complex deviants in schizophrenia. *Schizophrenia Research*, 191, 25–34. https://doi.org/10.1016/j.schres.2017.07.009

Biagianti, B., Roach, B. J., Fisher, M., Loewy, R., Ford, J. M., Vinogradov, S., & Mathalon, D. H. (2017). Trait aspects of auditory mismatch negativity predict response to auditory training in individuals with early illness schizophrenia. *Neuropsychiatr Electrophysiol*, 3, 2. https://doi.org/10.1186/s40810-017-0024-9

Bodatsch, M., Ruhrmann, S., Wagner, M., Muller, R., Schultze-Lutter, F., Frommann, I., ... Brockhaus-Dumke, A. (2011). Prediction of psychosis by mismatch negativity. *Biological Psychiatry*, 69(10), 959–966. https://doi.org/10.1016/j.biopsych.2010.09.057

Brown, W. (1910). Some experimental results in the correlation of mental Abilities1. *British Journal of Psychology, 1904–1920*, 3(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Chen, C., Chan, C. W., & Cheng, Y. (2018). Test-retest reliability of mismatch negativity (MMN) to emotional voices. *Frontiers in Human Neuroscience*, 12, 453. https://doi.org/10.3389/fnhum.2018.00453

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 86 (2), 127–137.

Cleveland, W. S., Grosse, E., & Shyu, W. (1992). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 309–376). New York: Chapman & Hall.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134 (1), 9–21.

Deouell, L. Y., & Bentin, S. (1998). Variable cerebral responses to equally distinct deviance in four auditory dimensions: A mismatch negativity study. *Psychophysiology*, 35(6), 745–754.

Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The global assessment scale. A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry*, 33(6), 766–771. https://doi.org/10.1001/archpsyc.1976.01770060086012

Erickson, M. A., Ruffle, A., & Gold, J. M. (2016). A meta-analysis of mismatch negativity in schizophrenia: From clinical risk to disease specificity and progression. *Biological Psychiatry*, 79(12), 980–987. https://doi.org/10.1016/j.biopsych.2015.08.025

Fryer, S. L., Roach, B. J., Ford, J. M., Donaldson, K. R., Calhoun, V. D., Pearlson, G. D., ... Mathalon, D. H. (2018). Should I stay or should I go? FMRI study of response inhibition in early illness schizophrenia and risk for psychosis. *Schizophr Bull*, 45, 158–168. https://doi.org/10.1093/schbul/sbx198

Fryer, S. L., Roach, B. J., Wiley, K., Loewy, R. L., Ford, J. M., & Mathalon, D. H. (2016). Reduced amplitude of low-frequency brain oscillations in the psychosis risk syndrome and early illness schizophrenia. *Neuropsychopharmacology*, 41(9), 2388–2398. https://doi.org/10.1038/npp.2016.51

Fryer, S. L., Woods, S. W., Kiehl, K. A., Calhoun, V. D., Pearlson, G. D., Roach, B. J., ... Mathalon, D. H. (2013). Deficient suppression of default mode regions during working memory in individuals with early psychosis and at clinical high-risk for psychosis. *Frontiers in Psychiatry*, 4, 92. https://doi.org/10.3389/fpsyt.2013.00092

Fulham, W. R., Michie, P. T., Ward, P. B., Rasser, P. E., Todd, J., Johnston, P. J., ... Schall, U. (2014). Mismatch negativity in recent-onset and chronic schizophrenia: A current source density analysis. *PLoS One*, 9(6), e100221. https://doi.org/10.1371/journal.pone.0100221

Glover, G. H., Mueller, B. A., Turner, J. A., van Erp, T. G., Liu, T. T., Greve, D. N., ... Potkin, S. G. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *Journal of Magnetic Resonance Imaging*, 36(1), 39–54. https://doi.org/10.1002/jmri.23572

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, 55(4), 468–484.

Hall, M. H., Schulze, K., Rijsdijk, F., Picchioni, M., Ettinger, U., Bramon, E., ... Sham, P. (2006). Heritability and reliability of P300, P50 and duration mismatch negativity. *Behavior Genetics*, 36(6), 845–857. https://doi.org/10.1007/s10519-006-9091-6

Hamilton, H. K., Roach, B. J., Bachman, P. M., Belger, A., Carrion, R. E., Duncan, E., ... Mathalon, D. H. (2019). Association between P300 responses to auditory oddball stimuli and clinical outcomes in the psychosis risk syndrome. *JAMA Psychiatry*, 76, 1187. https://doi.org/10.1001/jamapsychiatry.2019.2135

Hamilton, H. K., Woods, S. W., Roach, B. J., Llerena, K., McGlashan, T. H., Srihari, V. H., ... Mathalon, D. H. (2019). Auditory and visual oddball stimulus processing deficits in schizophrenia and the psychosis risk syndrome: Forecasting psychosis risk with P300. *Schizophrenia Bulletin*, 45(5), 1068–1080. https://doi.org/10.1093/schbul/sby167

Heyes, M. P., Ellis, R. J., Ryan, L., Childers, M. E., Grant, I., Wolfson, T., ... HNRC Group. (2001). Elevated cerebrospinal fluid quinolinic acid levels are associated with region-specific cerebral volume loss in HIV infection. *Brain*, 124(Pt 5), 1033–1042.

Jahshan, C., Cadenhead, K. S., Rissling, A. J., Kirihara, K., Braff, D. L., & Light, G. A. (2012). Automatic sensory information processing abnormalities across the illness course of schizophrenia. *Psychological Medicine*, *42*(1), 85–97. https://doi.org/10.1017/S003329171100 1061

Jernigan, T. L., Press, G. A., & Hesselink, J. R. (1990). Methods for measuring brain morphologic features on magnetic resonance images. Validation and normal aging. *Archives of Neurology*, *47*(1), 27–32.

Jones, S. H., Thornicroft, G., Coffey, M., & Dunn, G. (1995). A brief mental health outcome scale-reliability and validity of the global assessment of functioning (GAF). *The British Journal of Psychiatry*, *166*(5), 654–659.

Kathmann, N., Frodl-Bauch, T., & Hegerl, U. (1999). Stability of the mismatch negativity under different stimulus and attention conditions. *Clinical Neurophysiology*, *110*(2), 317–323.

Kiang, M., Braff, D. L., Sprock, J., & Light, G. A. (2009). The relationship between preattentive sensory processing deficits and age in schizophrenia patients. *Clinical Neurophysiology*, *120*(11), 1949–1957. https://doi.org/10.1016/j.clinph.2009.08.019

Koshiyama, D., Kirihara, K., Tada, M., Nagai, T., Fujioka, M., Koike, S., … Kasai, K. (2018). Association between mismatch negativity and global functioning is specific to duration deviance in early stages of psychosis. *Schizophrenia Research*, *195*, 378–384. https://doi.org/10.1016/j.schres.2017.09.045

Kujala, T., Kallio, J., Tervaniemi, M., & Naatanen, R. (2001). The mismatch negativity as an index of temporal processing in audition. *Clinical Neurophysiology*, *112*(9), 1712–1719.

Lew, H. L., Gray, M., & Poole, J. H. (2007). Temporal stability of auditory event-related potentials in healthy individuals and patients with traumatic brain injury. *Journal of Clinical Neurophysiology*, *24*(5), 392–397. https://doi.org/10.1097/WNP.0b013e31814a56e3

Light, G. A., & Braff, D. L. (2005). Stability of mismatch negativity deficits and their relationship to functional impairments in chronic schizophrenia. *The American Journal of Psychiatry*, *162*(9), 1741–1743. https://doi.org/10.1176/appi.ajp.162.9.1741

Light, G. A., Swerdlow, N. R., Rissling, A. J., Radant, A., Sugar, C. A., Sprock, J., … Braff, D. L. (2012). Characterization of neurophysiologic and neurocognitive biomarkers for use in genomic and clinical outcome studies of schizophrenia. *PLoS One*, *7*(7), e39434. https://doi.org/10.1371/journal.pone.0039434

Light, G. A., Swerdlow, N. R., Thomas, M. L., Calkins, M. E., Green, M. F., Greenwood, T. A., … Turetsky, B. I. (2015). Validation of mismatch negativity and P3a for use in multi-site studies of schizophrenia: Characterization of demographic, clinical, cognitive, and functional correlates in COGS-2. *Schizophrenia Research*, *163*(1-3), 63–72. https://doi.org/10.1016/j.schres.2014.09.042

Mathalon, D. H., Roach, B. J., Ferri, J. M., Loewy, R. L., Stuart, B. K., Perez, V. B., … Ford, J. M. (2018). Deficient auditory predictive coding during vocalization in the psychosis risk syndrome and in early illness schizophrenia: The final expanded sample. *Psychological Medicine*, *49*, 1897–1904. https://doi.org/10.1017/S0033291718002659

Mathews, C. A., Perez, V. B., Roach, B. J., Fekri, S., Vigil, O., Kupferman, E., & Mathalon, D. H. (2016). Error-related brain activity dissociates hoarding disorder from obsessive-compulsive disorder. *Psychological Medicine*, *46*(2), 367–379. https://doi.org/10.1017/S0033291715001889

McCleery, A., Mathalon, D. H., Wynn, J. K., Roach, B. J., Hellemann, G. S., Marder, S. R., & Green, M. F. (2019). Parsing components of auditory predictive coding in schizophrenia using a roving standard mismatch negativity paradigm. *Psychological Medicine*, *49*, 1–12. https://doi.org/10.1017/S0033291718004087

Naatanen, R. (2003). Mismatch negativity: Clinical research and possible applications. *International Journal of Psychophysiology*, *48*(2), 179–188.

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Pekkonen, E., Rinne, T., & Naatanen, R. (1995). Variability and replicability of the mismatch negativity. *Electroencephalography and Clinical Neurophysiology*, *96*(6), 546–554.

Perez, V. B., Ford, J. M., Roach, B. J., Loewy, R. L., Stuart, B. K., Vinogradov, S., & Mathalon, D. H. (2012). Auditory cortex responsiveness during talking and listening: Early illness schizophrenia and patients at clinical high-risk for psychosis. *Schizophrenia Bulletin*, *38*(6), 1216–1224. https://doi.org/10.1093/schbul/sbr124

Perez, V. B., Ford, J. M., Roach, B. J., Woods, S. W., McGlashan, T. H., Srihari, V. H., … Mathalon, D. H. (2012). Error monitoring dysfunction across the illness course of schizophrenia. *Journal of Abnormal Psychology*, *121*(2), 372–387. https://doi.org/10.1037/a0025487

Perez, V. B., Woods, S. W., Roach, B. J., Ford, J. M., McGlashan, T. H., Srihari, V. H., & Mathalon, D. H. (2014). Automatic auditory processing deficits in schizophrenia and clinical high-risk patients: Forecasting psychosis risk with mismatch negativity. *Biological Psychiatry*, *75*(6), 459–469. https://doi.org/10.1016/j.biopsych.2013.07.038

Pfefferbaum, A., Lim, K. O., Zipursky, R. B., Mathalon, D. H., Rosenbloom, M. J., Lane, B., … Sullivan, E. V. (1992). Brain gray and white matter volume loss accelerates with aging in chronic alcoholics: A quantitative MRI study. *Alcoholism, Clinical and Experimental Research*, *16*(6), 1078–1089.

Qiao, Z., Yang, A., Qiu, X., Yang, X., Zhang, C., Zhu, X., … Yang, Y. (2015). Gender effect on pre-attentive change detection in major depressive disorder patients revealed by auditory MMN. *Psychiatry Research*, *234*(1), 7–14. https://doi.org/10.1016/j.pscychresns.2015.05.011

Rahne, T., von Specht, H., & Muhler, R. (2008). Sorted averaging–application to auditory event-related responses. *Journal of Neuroscience Methods*, *172*(1), 74–78. https://doi.org/10.1016/j.jneumeth.2008.04.006

Recasens, M., & Uhlhaas, P. J. (2017). Test-retest reliability of the magnetic mismatch negativity response to sound duration and omission deviants. *NeuroImage*, *157*, 184–195. https://doi.org/10.1016/j.neuroimage.2017.05.064

Roach, B., Carrion, R. E., Hamilton, H. K., Bachman, P., Belger, A., Duncan, E., Jason Johannesen, Gregory A. Light, Margaret Niznikiewicz, Jean Addington, Carrie E. Bearden, Kristin S. Cadenhead, Tyrone D. Cannon, Barbara A. Cornblatt, Thomas H. McGlashan, Diana O. Perkins, Larry Seidman, Ming Tsuang, Elaine F. Walker, Scott W. Woods Mathalon, D. H. (2019). Reliability of mismatch negativity event-related potentials in a multisite, traveling subjects study. doi: https://www.biorxiv.org/content/10.1101/768408v1

Sakamoto, Y., & Kitagawa, G. (1987). *Akaike information criterion statistics*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Schall, U., Catts, S. V., Karayanidis, F., & Ward, P. B. (1999). 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. *The International Journal of Neuropsychopharmacology*, *2*(2), 83–93. https://doi.org/10.1017/S1461145799001418

Schroger, E., Giard, M. H., & Wolff, C. (2000). Auditory distraction: Event-related potential and behavioral indices. *Clinical Neurophysiology*, *111*(8), 1450–1460.

Shaikh, M., Valmaggia, L., Broome, M. R., Dutt, A., Lappin, J., Day, F., … Bramon, E. (2012). Reduced mismatch negativity predates the onset of psychosis. *Schizophrenia Research*, *134*(1), 42–48. https://doi.org/10.1016/j.schres.2011.09.022

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Tervaniemi, M., Lehtokoski, A., Sinkkonen, J., Virtanen, J., Ilmoniemi, R. J., & Naatanen, R. (1999). Test-retest reliability of mismatch negativity for duration, frequency and intensity changes. *Clinical Neurophysiology*, *110*(8), 1388–1393.

Todd, J., Michie, P. T., Schall, U., Karayanidis, F., Yabe, H., & Naatanen, R. (2008). Deviant matters: Duration, frequency, and intensity deviants

reveal different patterns of mismatch negativity reduction in early and late schizophrenia. *Biological Psychiatry*, *63*(1), 58–64. https://doi.org/10.1016/j.biopsych.2007.02.016

Uwer, R., & von Suchodoletz, W. (2000). Stability of mismatch negativities in children. *Clinical Neurophysiology*, *111*(1), 45–52.

Witkovský, V. (2012). Estimation, testing, and prediction regions of the fixed and random effects by solving the Henderson's mixed model equations. In *Measurement science review* (Vol. 12, p. 234).

Yang, X., Yu, Y., Chen, L., Sun, H., Qiao, Z., Qiu, X., … Yang, Y. (2016). Gender differences in pre-attentive change detection for visual but not auditory stimuli. *Clinical Neurophysiology*, *127*(1), 431–441. Warsaw Poland: Sciendo. https://doi.org/10.1016/j.clinph.2015.05.013

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.